Structure-Based Methods for Virtual Screening, Protein Design, and Protein Function Studies

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Pietro Alfarano

aus
Italien

Promotionskomitee

Prof. Dr. Amedeo Caflisch (Leitung der Dissertation)
PD Dr. Peer Mittl
Prof. Dr. Cristina Nevado

Zürich, 2010

Per Sofie

Contents

List of Figures

Summary

The average time a drug requires to successfully appear into the market is around 15–20 years. Structure-based techniques play a determinant role in many moments of this long process and computer-based approaches are becoming always more popular, for they are cheap and fast. Computer-based approaches to structure-based drug discovery are also very common in Academia, both in the field of development and application. Moreover, most of the software currently used in this field is being developed in Academia.

In this doctoral thesis two main topics are covered: virtual screening by the means of docking and the application of molecular dynamics simulations to macromolecular systems of interest.

Two projects employed virtual screening: the search for inhibitors of cathepsin B and the search for allosteric effectors of cathepsin K. The two projects are ideally linked because in both the active site of the enzyme was not targeted. In the cathepsin B project, the so called "occluding loop" was chosen to show that inhibition of this enzyme could be obtained without targeting the active site. In the cathepsin K project, a putative allosteric pocket was targeted. The pocket was found by a method that considers the residue conservation in evolutionarily linked protein families and then confirmed by two independent structure-based methods.

Two other projects involved molecular dynamics simulations of biological macromolecules: designed armadillo repeat proteins and the CRY1 protein of Arabidopsis. In both projects, the evaluation of local and conformational flexibility played a major role. The first one employed static homology modelling and molecular dynamics simulations not only to generate a putative structure of an armadillo repeat protein, but also for improving its dynamical behavior, which could not have been inferred from a static structure only. The second project helped to explain the effect of a single-point mutation of the CRY1 protein of Arabidopsis thaliana. Molecular dynamics simulations of the two systems were run and the difference of activity between them was explained in terms of different flexibility of the protein surface close to the mutation.

Zusammenfassung

Die durchschnittliche Zeit, die ein Medikament braucht, um erfolgreich auf den Markt zu kommen, beträgt 15-20 Jahre. Struktur-basierte Methoden spielen eine entscheidende Rolle an vielen Stellen dieses langen Prozesses und computer-basierte Ansätze werden immer beliebter, da sie kostengünstig und schnell sind. Computer-basierte Ansätze für die struktur-basierte Identifikation neuer Medikamente werden sehr häufig in der Wissenschaft verwendet, sowohl auf dem Gebiet der Entwicklung als auch für Anwendungen. Darüber hinaus wird ein Grossteil der aktuellen Software dieses Gebiets in der Wissenschaft entwickelt.

In dieser Doktorarbeit werden zwei Hauptthemen behandelt: virtuelles Screening durch Docken und die Anwendung von Moleküldynamik-Simulationen auf makromolekulare Systeme.

Bei zwei Projekten wurde virtuelles Screening verwendet: die Suche nach Inhibitoren von cathepsin B und die Suche nach allosterischen Effektoren von cathepsin K. Die zwei Projekte sind ideal miteinander verbunden, da beide nicht auf das aktive Zentrum der Enzyme abzielten. Im cathepsin B Projekt wurde die sogenannte "occluding loop" gewählt um zu zeigen, dass das Enzym inhibiert werden kann, ohne das aktive Zentrum des Enzyms anzugreifen. Im cathepsin K Projekt wurde eine mutmassliche allosterische Bindungsstelle ausgewählt. Die Bindungstasche wurde mit Hilfe einer Methode gefunden, die den Erhalt der Aminosäuresequenz bei evolutionär verbundenen Proteinfamilien berücksichtigt, und anschliessend durch zwei unabhängige struktur-basierte Verfahren bestätigt.

Zwei weitere Projekte beinhalten Moleküldynamik-Simulationen von biologischen Makromolekülen: entworfene Armadillo Repeat-Proteine und das CRY1 Protein der Arabidopsis. In beiden Projekten spielte die Evaluation der lokalen und konformationellen Flexibilität eine zentrale Rolle. Beim ersten wurden statisches Homologie-Modelling und Moleküldynamik-Simulationen verwendet, um nicht nur eine mutmassliche Struktur eines Armadillo Wiederholungsproteins zu generieren, sondern auch sein dynamisches Verhalten zu verbessern, das allein aus einer statischen Struktur nicht

hätte abgeleitet werden können. Das zweite Projekt diente der Erklärung der Auswirkung einer Einzelmutation des CRY1 Proteins von Arabidopsis thaliana. Moleküldynamik-Simulationen beider Systeme (Wildtype und Mutierter) wurden durchgeführt und der Unterschied ihren Aktivitäten mit einer unterschiedlichen Flexibilität der Proteinoberfläche in örtlicher Nähe zur Mutation erklärt.

Chapter 1

Introduction

In the introduction of the thesis several topics will be covered: the drug discovery and development process, structure-based drug design (docking, de-novo design, and scoring), and molecular dynamics simulations of macromolecules. These topics are strongly interconnected and knowing their applicability and possibly also limitations is fundamental to the understanding of how drugs are developed.

In 1913, the German scientist Paul Ehrlich coined the Latin phrase *"corpora non agunt nisi fixata"*, that can be literally translated to "bodies do not act, unless they are bound", and poetically to "molecules are not biologically active, unless they are bound to one or more receptors and therefore trigger or block some biological response", and most of the efforts of the computational methods presented in the thesis try to predict whether a drug binds to an enzyme (docking and de-novo design), and, if it happens, how strong is the interaction (scoring).

Moreover, since proteins and drugs are not fixed entities, as crystal structures suggest us, molecular dynamics simulation is another important aspect of drug design. Section 1.9 briefly explains the basis of molecular dynamics.

1.1 The drug discovery process

Drug discovery can be provocatively described as "an expensive and time-consuming activity that mostly fails"[42]. Retrospective analyzes of the pharmaceutical industry market performed during the nineties estimated that new drugs needed 14 years in average to appear into the market, each one costing about 800 million dollars. Additionally, one compound every nine that enters clinical trials eventually hits the market. Even if the

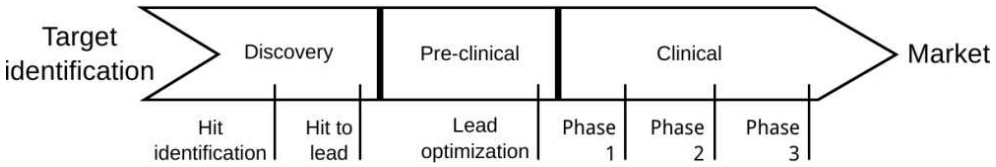


Figure 1.1: From target identification to the market. The figure schematically represents the phases of the development of a new drug entity. Redrawn from [42].

aforementioned scenario could have changed a bit during the last years, a meaningful message remains: drug discovery is a tough activity and wrong decisions taken at the initial steps of the development often have a very deleterious influence on the success of the drug candidate. It's common for pharmaceutical industries to have more than one compound in development, also to cover and minimize the possible failure of a drug candidate. Not to mention that, if the marketed drug is a blockbuster, the revenues will be impressive.

A general scheme of the process of drug discovery is depicted in figure 1.1. The phases are the usual, accepted way to describe the process, but the directions and divisions are not always so neat as it appears from the scheme. Nevertheless, the process is here presented in several separated moments[42], as it follows.

1.1.1 Target identification

Naturally, the first step is the identification of a relevant protein target such as enzymes or receptors. The identification of a target necessarily requires some level of validation and the starting point is establishing a biological and biochemical rationale for why affecting a particular target will have the desired therapeutic benefit. Establishing a rationale is particularly necessary for structure-based drug design.

1.1.2 Hit identification

A compound that binds to the target and shows the desired biological effect is called "hit". The most common way to find an initial hit is by screening a compound library which can consist of natural products or substrate analogues/mimetics, an *in-house* library of compound, a commercial library,

or a targeted library[1]. Compounds can be screened experimentally, for example with High Throughput Screening (HTS), or computationally (Virtual Screening, by means of docking, for example). The outcome of HTS is a pool of compounds whose activity is also checked for reproducibility in experimental conditions different from the HTS ones, and whose chemical structure and purity have been checked. Depending on the size of the screened library, hit identification requires typically 6–12 months and significant resources.

1.1.3 Hits to leads

This is one of the most delicate phases of drug development. It is usual to re-synthesize the hits for a complete validation and often dramatic changes in the chemical template are made for establishing the essential core of the lead compound. Structure Activity Relationships (SAR[2]) are established, as well as physio-chemical and ADMET[3] properties, chemical tractability, synthetic accessibility, and the Intellectual Property (IP) position on the hit compound and the derived series. Setting the right requirements in this phase is crucial for the lead optimization and it is one of the most challenging aspects of medicinal chemistry. This phase usually lasts 6 months, even if complicate syntheses and biological assays can definitely elongate it.

1.1.4 Lead optimization

The main goal of this phase, the most resource-intensive in drug discovery, is to develop one or more compounds with particular properties. Affinity for a target and selectivity are desired properties as well as drug-like properties and cell permeability. Selectivity requirements can be less stringent

[1] A targeted library is a library of compounds that complies with some criteria, such as particular hydrogen bonding patterns. For instance, kinases have mostly one conserved hydrogen bond donor and one acceptor in the so called *hinge region*. Several kinase inhibitors have one hydrogen bond acceptor and one donor in the correct geometry for satisfying optimal bonding pattern.

[2] Structure Activity Relationship is a very important moment of drug discovery in which medicinal chemistry plays a significant role. Several modifications to the hit are introduced to study their impact on the compound activity. The modifications can be introduced, for example, by analogy to a known active compound; by a rationale based on the structure of the complex between the biological target and the hit, also obtained by computational means like docking; or by the modifications of functional groups with isosteres.

[3] ADMET is an acronym that stands for Absorption, Distribution, Metabolism, Excretion and Toxicity. All these processes are what the body "does" to the drug and pharmacokinetics (PK) is the discipline that analyzes those aspects.

if the drug is devoted to an acute condition (such as cancer, where the adverse effects related problems are less severe than the condition). The main changes introduced in the compound at this point of development are not in the "center" of the molecule, but rather in the "periphery". Remarkably, very small and subtle changes can have a very dramatic effects when tested in cells or *in vivo* (for example, inserting a single methyl on a phenyl ring can introduce a reactive site for cytochrome oxidation, therefore increasing the *clearance* by excretion). This phase can last 18–36 months. After lead optimization, one or more compounds are available with the desired criteria of efficacy on animal models, with a demonstrable mode of action and acceptable pharmacokinetics (ADMET).

1.1.5 Pre-clinical trials

This phase is setting up and preparing the compounds for testing in humans. This includes scale-up synthesis (a synthesis effective on the milligrams scale might not be successful on the hundreds of grams scale), formulation, toxicology, and design of the clinical trials. Most of this phase is covered by a stringent regulatory regime and it is strongly required to work according to certain legal guidelines.

1.1.6 Clinical trials

This is the most expensive and time consuming part of the development of a new drug. It is conventionally separated in three stages.

Phase 1. Assessment of drug candidate's safety. A small group of healthy volunteers are given the compound and the pharmacokinetics properties of the drug are assessed. This might take up to several months. About 70% of the drugs pass this phase.

Phase 2. Assessment of drug efficacy. A bigger group of persons affected by the condition to be treated takes part to a randomized double-blind experiment[4], where the drug is mainly tested for efficacy and for safety. About only 30% of the projects successfully complete phase 1 and 2. After the completion of phase 2, a compound can be considered a drug. Phase 2 can last from several months to years.

Phase 3. Assessment of effectiveness, benefits, and possible adverse reactions. The drug is administered from several hundreds to several thousands

[4] A double-blind test is a way of eliminating the subjective bias on the patient and the experimenter, whose outcome can be the *placebo effect*. Shortly, in a double-blind test, neither the patient, nor the experimenter know whether the (experimental) drug or a placebo is dispensed.

of patients, and this phase usually lasts many years. It's usual to compare the drug effectiveness with existing treatments on the market for assessing increased benefits. These trials provide the necessary data for obtaining approval by the regulatory authorities.
Even after the drug appears into the market, continued trials and monitoring are required (the *phase 4*). Some rare adverse effects can appear only when a drug is given to many people. Around 20 drugs have been withdrawn from the market for this reason in the last 10 years[5].

1.2 Structure-based drug discovery

Computational structure-based drug discovery plays a fundamental role in medicinal chemistry. In two recent reviews[46, 52], the authors notice that no drug has been hitherto discovered entirely by structure-based computational methods, mostly because these methods contribute essentially to the early phases of drug-discovery, such as lead finding and lead optimization. Interestingly, the number of projects for which structure-based drug discovery has played fundamental a key role is around 10.

1.2.1 *Non*-structure-based drug discovery

When the structure of the target is unfortunately unknown, the previously obtained Structure Activity Relationship data (SAR) can be used, for example, to build up a pharmacophore model[6]. The compound series employed for the SAR study is superimposed and the geometry and the spatial constrains of the pharmacophoric groups are then screened in compounds libraries, assuming that the computer-generated conformation is very similar to the bio-active one. Additionally, active compounds can be found by similarity searching using known binders as templates, assuming that molecules

[5]Famous examples of drugs withdrawn from the market are *cerivastatin*/Lipobay (2001) and *rofecoxib*/Vioxx (2004). The first was a drug used to lower cholesterol levels in blood, thus preventing cardiovascular diseases. It was withdrawn because it caused *rhabdomyolysis*, a rapid breakdown of skeletal muscle, which can lead to kidney failure because of the renal toxicity of myoglobin, an oxygen-binding protein found in the muscle tissue. The second drug was a non-steroidal anti-inflammatory drug, which was withdrawn because of increased risk of hearth attack and stroke associated with long-term, high-dosage use.

[6]The IUPAC definition of pharmacophore is "an ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target and to trigger (or block) its biological response model for the positioning of key features like hydrogen-bonding, charged and hydrophobic".

with the similar shape bind in the same fashion (even if sometimes it is not the case[5]). The similarity can be a simple scaffold similarity[7] or a more complicated search for molecules that share in the same volume the same steric and electronic properties (a very good program for this task is ROCS, by OpenEye, Inc.).
However, the aim of this introduction is to illustrate concisely the structure-based drug discovery methods, such as docking and de-novo design.

1.2.2 Basic concepts

Some basic concepts for the understanding of the thesis are succinctly introduced. The terminology employed in the field refers to the small molecule as **ligand** or **binder** and to the protein, to which the ligand binds, as **receptor**. In structure-based drug design, the chief experimental source for obtaining structures is X-ray crystallography. It is possible by several experimental means to obtain the cocrystallization of a protein and a bound small ligand, often an inhibitor. A particular receptor-bound conformation of a ligand is called **pose** or **binding mode**. Moreover, the binding geometry found in the experimentally determined structure is often called **native pose**, **native binding mode**, or **X-ray pose**. A **binding pocket** is a part of the (solvent accessible) surface of a macromolecule to which usually a (small) molecule binds. It can be defined from the position of a cocrystallized compound in an X-ray structure (see figure 1.2), or by geometric analysis[56], if the structure is devoid of such a binder. The **active site** of an enzyme is the portion of space in which the enzymatic activity is performed. Usually the largest binding pocket of an enzyme is also its active site. Other auxiliary pockets near the active site define the **specificity pockets** or **sub-pockets**, which are exploited by natural or artificial ligands for increasing their affinity and selectivity to the receptor.

A common and computationally efficient approximation introduced for the calculation of interaction energies is the **pair-wise** additive approximation. According to it, the interaction energy of one atom with the rest of the system is calculated as the sum of this atom with all the others in a pair-wise manner. This approximation does not take into account the simultaneous interaction between three or more atoms, such as polarization effects.

[7] For example, if it is known that the important pharmacophore is a pyrimidine and a morpholine, one can screen for those moieties in bigger libraries.

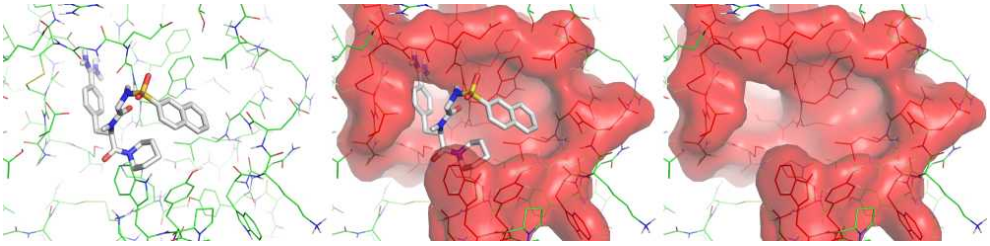


Figure 1.2: Binding pocket definition. In the docking procedure developed in our group (see section 1.7) the binding pocket (red surface) is usually defined from a cocrystallized compound (white carbons, stick representation) of an X-ray structure and consists of residues which are in close contact (distance ≤ 5 Å) with the aforementioned compound.

1.3 Ligand docking, issues and challenges

Since 1971 to nowadays, the number of public available protein crystal structures has increased from 7 to more than 55000[8] and thousands new structures are being added every year. This massive amount of structural information is also flanked by several collections of databases of purchasable compounds. For example, the freely accessible ZINC database[44] contains more than 13 million compounds[9].

Since 1982, when Dock[55], the first small molecule to protein docking program, was described, many new and different programs have been developed and the most used in academia and industry are: Dock, Fred, Glide[26], Gold[45], FlexX[67], and AutoDock[32], just to name a few.

Docking programs mainly differ for the computational algorithm employed for finding the bioactive conformation (exploration of the conformational space), for placing it correctly in the binding pocket, and for calculating the interaction energy between the pose and the receptor. In other words, docking is solving the molecular recognition problem in biological systems. A recent review by Stahl[4] covers the main interactions responsible of ligand binding and molecular recognition. The calculation of the interaction energy will be discussed more in detail in section 1.5 by illustrating the scoring process. Docking programs are designed for being able to identify, reproduce and optimize specific attractive interactions between the two partners. Requirements of docking software are speed and reliability, and usually speed is achieved at the expense of reliability. Speed

[8] http://www.rcsb.org/pdb/statistics/contentGrowthChart.do accessed on April 6, 2010.

[9] http://zinc.docking.org

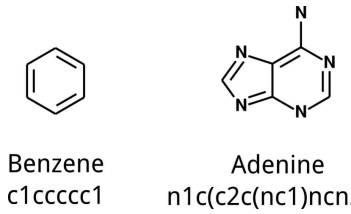


Figure 1.3: SMILES. Benzene and adenine as 2D and SMILES notation.

is required for the efficient virtual screening of large libraries of compounds. Section 1.7 provides an example of a docking software.

Normally, the 3D X-ray conformation of the binder is not fed to the docking program and ligand conformations are automatically generated from a 1D representation like SMILES[80] (see figure 1.3 for an example of SMILES representation) with an energy-minimized geometry, and the most appropriate tautomeric form and protonation state.

1.3.1 Correctness of docking

An RMSD (Root Mean Square Deviation, see section 1.9.4) between the native X-ray pose and the docked poses of lower than 2 Å is generally considered a success. An RMSD of 2 Å is not too permissive, even if, in a limit case, it can be interpreted as a rigid translation of 2 Å of the molecule from its original position. In fact, unless the resolution of the crystal structure is very high, there is an inherent uncertainty in the position of the native pose. Additionally, docking programs are generally unable to reproduce deviations of bond distances and angles from the equilibrium value, as sometimes observed in native conformations.

1.3.2 Importance of the tautomeric form

The tautomeric form[59] and the protonation state of the ligand and the receptor can play a determinant role in binding[74]. A molecule exhibits tautomerism if it is representable by two or more structures, which are connected by the movement of an hydrogen from one atom to another of the same molecule. The prediction of the most appropriate tautomeric form is of utter importance because each tautomer of a single molecule substantially differs from each other in electrostatic properties, hydrophobicity, tree dimensional shape, and chemical reactivity. For example, the proton shift

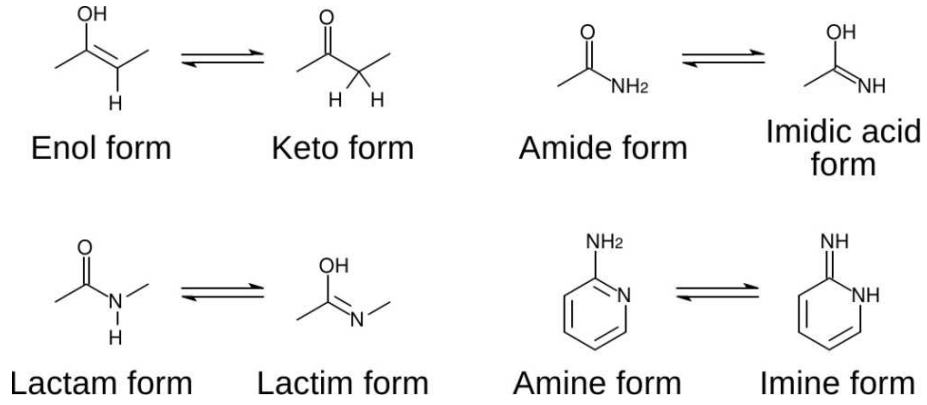


Figure 1.4: Tautomerism. Four frequent examples of tautomerism in medicinal chemistry are shown.

from the enol-form to a keto-form (see figure 1.4) of a molecule changes the alcohol group into a carbonyl group, hence transforming a hydrogen bond donor to an acceptor and this inversion of reactivity can have a determinant role in ligand binding. The importance of tautomers is not only restricted to computational efforts. A recent analysis of the frequency of marketed drugs that can tautomerize[59] showed that of the 1791 compounds, 1334 or 74% exist as only one tautomer. The remaining 26% have an average of three tautomers.

1.3.3 Importance of the protonation state

If a particular chemical group of the ligand or the receptor is neutral or charged, the erroneous assignment of the formal charge through the protonation state can also have severe effects on the correctness binding. The protonation state plays a very important role for proteins, too. In fact, at physiological pH (circa 7.4), it is generally assumed that the guanidine group of the side chain of arginine (pK_a=12.48[10]) and the ϵ-amino group of lysine (pK_a=10.54) are protonated (positively charged), while the carboxy groups of the side chains of aspartic and glutammic acids (pK_a=3.39 and pK_a=4.07, respectively) are not protonated (negatively charged). The properties of the histidine side chains (pK_a=6.04, imidazole) are even more difficult to predict. Additionally, the exact orientation of the N and O atoms

[10]The pK_a values have been taken from "Fundamentals of Biochemistry", by D. Voet, J. G. Voet, and C. W. Pratt, John Wiley and Sons, 1999. — A very kind present of Prof. Antonio Baici to the author of this doctoral thesis.

of asparagine and glutamine, and of the imidazole ring of histidine, is virtually indistinguishable, except for protein structures at very high resolution. The same problem arises for the position of hydrogens bound to aspartic and glutammic acids, and to the C-terminus (if the pH is acidic enough), and to the N_δ or N_ϵ of histidine side chains. One solution[11] to orient those residues according to hydrogen bond networks[17]. Widely used software for the enumeration of tautomeric forms and the prediction of pK_a values are: LigPrep and Epik (Scrödinger, LLC), and QuacPac (OpenEye, Inc)[12].

Moreover, it is generally assumed that, upon binding, there is no change in the protonation or tautomeric states for the receptor and the ligand, but the exact protonation state strongly depends on the dielectric conditions imposed by the local environment. For example, an acidic ligand group which is placed in an apolar or in a positively charged protein environment will become less acid or more acid, respectively, compared to aqueous solution.

1.3.4 Re-docking, cross-docking, apo-docking, and the induced fit problem

During the docking process, docking programs frequently modify only torsional angles of a molecule, leaving bond distances and bond angles unaltered.

Given a cocrystallized receptor-ligand complex, a good docking program should be able to reproduce the X-ray binding mode of the ligand (the *re-docking* experiment), that is finding the bioactive conformation of the ligand and placing it correctly in the binding pocket.

If several X-ray structures of compounds cocrystallized with the same protein are available, a more interesting docking experiment is the *cross-docking* experiment (an exhaustive and insightful article on cross-docking by Verdonk is [77]). In this experiment, one tries to dock all the cocrystallized compounds in all the receptors structures[13] and it is not uncommon that this experiment fails to reproduce the native X-ray binding pose for some

[11] Several on-line servers are devoted to checking and validate crystal structures. One of the most used is MolProbity: http://molprobity.biochem.duke.edu/

[12] http://www.schrodinger.com and http://www.eyesopen.com, respectively.

[13] For example, if one has two crystal structures (A and B) of the same receptor with two different cocrystallized compounds (a and b), one docks the compound b in the receptor A and the compound a in the receptor B.

ligands. If the normal success rate for re-docking experiments is 70–100%[14], the success rate for cross-docking experiment is usually 40–50%.

In fact, upon binding, some side chains of the receptor (or even the backbone) can modify their torsional state to better accommodate the ligand (*induced fit*, see section 1.3.5), and the resulting conformation might be unfavorable to the binding of another compound. The problem is particularly important if the protein structure has been produced through homology modeling and the position of some side chains of the binding pocket is ambiguous. Moreover, if for the same receptor a holo-structure and a apo-structure are available, it is not uncommon that docking the co-crystallized compound to the apo-structure will fail (*apo-docking*) because some side chains in the apo-structures assume rotameric states which are unfavorable to the binding of the compound. Some docking protocols have been devised to cope with the problem of the *induced fit* with a soft docking strategy, that attenuates the steric clashes[69], or allowing different rotameric states in the side-chains[27] without allowing backbone movements. Moreover, it has been shown[64] that fixing the backbone, and allowing only side chains movements, can be responsible of docking failures. There are also approaches for including backbone mobility for improving docking accuracy: performing parallel docking using different conformations of the protein[11]; using computational techniques such as normal mode analysis to perturb a crystal structure into relevant structures suitable for docking, especially when loop movements are involved[12]; combine experimentally determined structures originating from multiple conformations[16].

Unfortunately, a common speed-up technique for docking, that is pre-calculating the pair-wise interactions in a dense 3D lattice comprising active site space, thus providing a *look-up table*, can not be used if several receptor conformations are employed. Side chains movements would modify the structure and make this approach too time-consuming, because one should recompute the look-up table for each new rotameric state. Recently, it has been shown that, using elastic potential grids[49], it is possible to adapt the pre-calculated 3D lattice to (small) variations in the binding pocket.

The problem of not keeping the protein fixed during docking, or allowing a protein structure minimization after docking, is that unfavorable repulsive interactions (locally introduced strain) can be dissipated to such an extent, by absorption by other parts of the structure, that they become unrecog-

[14] A success rate close to 100% for a particular target could have several reasons. The docking program could be the best one ever written or, more likely, the performance assessment could slightly biased because it is performed on targets belonging to the same protein families on which the docking program has been optimized.

nizable. It's therefore advisable to allow flexibility of only those side-chains that are known to adapt to ligand binding.

Finally, the role of conserved water molecules in crystal structures is gaining more importance[34]. The degree of conservation of the water molecules is hard to estimate and it is not clear how to classify tightly and loosely bound water molecules. Neglecting a tightly bound water molecule during docking can result in a high penalty for desolvation, therefore a unfavorable contribution to the binding affinity. The unfavorable contribution to the binding affinity is important when the water molecule makes more than two hydrogen bonds with the receptor[9]. In fact the hydroxyl group, which is the closest organic group to water, lacks the other hydrogen of water. Many docking protocols can easily include water molecules as a part of the receptor.

1.3.5 Induced fit or conformational selection?

Briefly, two different theoretical models describe the conformational changes of the receptor occurring upon ligand binding: induced fit and conformational selection.

Induced fit, formulated by Koshland in 1958[54], postulates that first a molecule binds to a receptor and then the conformational changes in the macromolecule follow.

On the contrary, conformational selection[75] postulates that a free enzyme exists in an equilibrium of different conformations sub-states and that the ligand preferentially binds to one or more sub-states, therefore shifting the equilibrium in such direction.

The conformational selection hypothesis is computationally appealing because it implies that it is possible to study a free enzyme with molecular dynamics simulations and then utilize some of the snapshots for docking. An example of the influence of conformational selection on docking is presented in section 1.9.5.

1.4 De-novo design

While docking tries to find the best arrangement of a particular molecule inside a binding pocket of an enzyme, the aim of de-novo design is conceptually different. During a de-novo design experiment, new molecules are generated "inside" the binding pocket, through the coupling of building blocks such as fragments or by growing algorithms. This task is generally very difficult because of the very huge search space involved. The gener-

ation of new molecular entities is often against their synthetic feasibility, and this aspect can reduce the usefulness of such programs. Several programs such as SMoG[18], SPROUT[29], LigBuilder[79], and GANDI[19] are available for de-novo design. Section 1.8 provides an example of a de-novo design software (GANDI).

1.5 Scoring

After a docking or a de-novo design experiment is performed, many poses are generated (see figure 1.5), and a *scoring function* is always used to distinguish docked or newly generated poses whose binding mode resembles a native binder from others. The scope of virtual screening is therefore to filter a set of molecules extracted from a database by discarding compounds which are likely to be inactive, prior to enzymatic tests, and keeping compounds which are likely to be active. Docking and de-novo algorithms implement energy functions, that are used to select energetically favorable poses, thus more likely to bind, but usually such energy functions are not sophisticated.

The problem of molecular recognition is very important (especially in biology) and scoring functions should be able to capture the physical principles responsible for binding. Importantly, scoring functions have to be sufficiently fast, efficient. For example the LIE method[1] for predicting binding energies is based molecular dynamics simulations and therefore requires several hours for a single pose. It is certainly not suitable for screening a big library of compounds.

Scoring functions have to provide the user with several requirements. First, the docked poses have to be ranked correctly. For example, in a redocking experiment, the pose with the most favorable score, thus ranked first, should be the closest (RMSD) to the native co-crystallized conformation. Second, the trend of the experimental binding energies should be reproduced, that is ligands should be ranked the same according to their score and to the experimental binding energy. Third, docking can be used as a virtual screening tool. The scoring functions should be able to select an active molecule from a pool of (many) inactive decoys[33]. Fourth, active and inactive compounds of a single chemotype should be distinguished. Often, after a lead compound is found, small chemical modifications are introduced to improve its potency. Scoring functions should be able to rank correctly the compound derived from a very similar scaffold. For example, more sophisticated and time demanding scoring functions can be able to distinguish the activity of closely related molecules, or for compound that

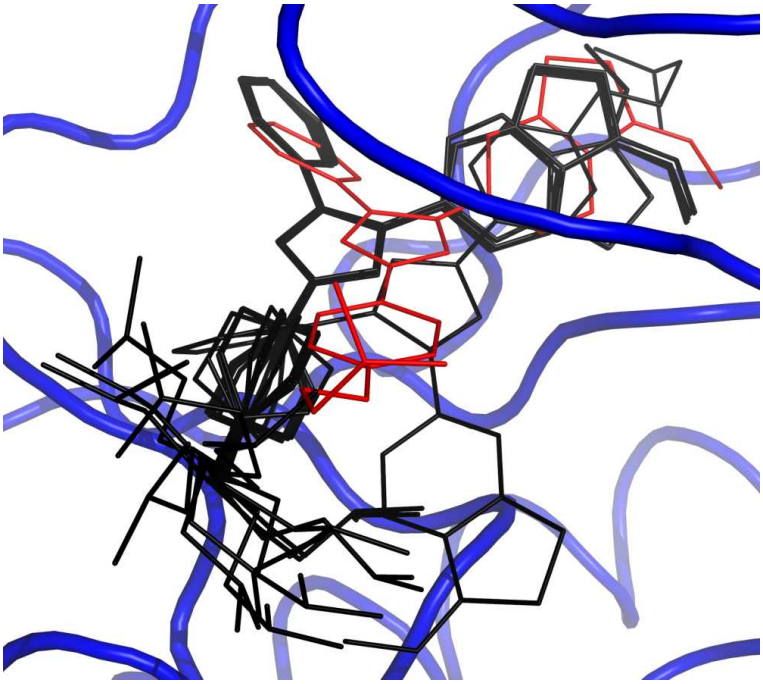


Figure 1.5: Typical results of docking. The receptor is in blue and the co-crystallized pose in red. The first 20 docked poses are in black. A very similar situation can be also obtained by de-novo design.

Figure 1.6: A hard test for scoring functions. Oxazole and isoxazole are two isomers and scoring functions sensitivity can be not high enough to apreciate the difference.

contain only slightly differing groups. In the oxyazole example (see figure 1.6), the distance of oxygen and nitrogen in the two isomers is too small for being apreciated by the sensitivity of simple scoring functions.

Scoring functions could be divided into three classes: force-field-based, empirical, and knowledge-based scoring functions. (For comprehensive reviews, see the reviews by Klebe[31] and by Stahl[70]).

It's important to notice that scoring functions usually correlate with molecular size[76] (see figure 1.7). For this reason, it is often better to compare compounds not on their calculated binding energy (or score), but

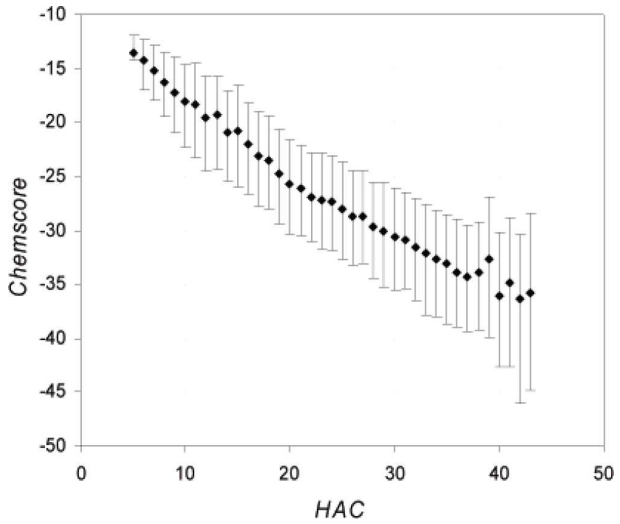


Figure 1.7: Scoring functions dependence on molecular size. A typical dependency of the Chemscore on the heavy atom count. It is a result of the docking of about 100 000 random compounds, with varying molecular weights. The error bars reflect the variance in the Chemscore values. Reproduced from [76].

on the binding efficiency, obtained by dividing their binding energy by their molecular weight or the heavy atom count (*ligand efficiency*).

1.5.1 Force-field based scoring functions

Force-field based scoring functions consist of the pair-wise non-bonded interaction terms (van der Waals and electrostatic contributions). For example, the CHARMM[7, 6] van der Waals interaction energy between the atom i and the atom j is approximated with a 12-6 Lennard-Jones potential:

$$\mathrm{V}_{LJ} = \epsilon_{i,j}\left(\left(\frac{\mathrm{R}_{min,i,j}}{r_{i,j}}\right)^{12} - 2\left(\frac{\mathrm{R}_{min,i,j}}{r_{i,j}}\right)^{6}\right)$$

where $r_{i,j}$ is the distance between the two atoms; $\epsilon_{i,j} = \sqrt{\epsilon_i \epsilon_j}$, ϵ_i and ϵ_j are constants characteristics of the strength of the van der Waals interaction of the two atoms; $\mathrm{R}_{min,i,j} = (R_{min,i} + R_{min,i})/2$, where $\mathrm{R}_{min,i}$ and $\mathrm{R}_{min,j}$ are constants characteristics of radii of the two atoms. The repulsive $1/r^{12}$ term dominates when the separation between the atoms is small. The $-1/r^6$ term is attractive and dominates when the separation of the two atoms increases. Alternatively, when *soft-docking* protocols are employed

to allow some minor steric clashes between the ligand and the receptor, the 12-6 Lennard-Jones potential is usually scaled to a more attenuated 8-4 one, where the repulsive term is less predominant. The electrostatic interaction term between two atoms is described in CHARMM by the Coulomb law:

$$\mathrm{V}_{elec} = \frac{q_i q_j}{4\pi\epsilon r}$$

where q_i and q_j are the partial charges of the atoms i and j, ϵ is the dielectric constant, and r is the separation between the two atoms. Force-field scoring functions were among the first to be introduced in docking and scoring, but they tend to overweight polar interactions. For example, when a salt-bridge is established, its calculated coulombic interaction energy is very favorable and often three or four times higher than the total van der Waals interaction, and this effect can generate artifacts. It has been shown in an assessment study of several scoring functions[24], that the influence of the protonation state of ligand and protein deeply affects the performance of force-field scoring functions, if the partial charges are not correctly assigned. However, taking into account the desolvation energies[71, 40] can compensate for this problem because salt-bridges have a high desolvation penalty which counterbalances the coulombic energy.

1.5.2 Empirical scoring functins

Empirical scoring functions, also known as regression-based scoring function, try to correlate the experimental ΔG of binding of protein complexes of determined structure to an empirical function, which is essentially a weighted sum of many terms which are thought to contribute to the free energy of binding[15]. The individual weights of the terms are determined by means of multiple linear regression. Usually, hydrophobic interactions and hydrogen bonding contributions are introduced, as well as entropy contributions. Two examples are the ChemScore[22] and the LIECE scoring functions[40]. The scoring function of ChemScore is based on the following terms: hydrogen bonds, metal interaction[16], lipophilic interactions and number of frozen rotatable bonds. The hydrogen bonds term is also scaled down for deviations from the most favorable geometry. The lipophilic term is calculated for all lipophilic atoms of the ligand and the receptor. Lipophilic atoms are chlorine, bromine and iodine (not in the ionic form),

[15] http://lpdb.chem.lsa.umich.edu/ and http://www.pdbbind.org are comprehensive collections of experimentally measured binding affinity data (K_d, K_i, and IC_{50}) for the protein-ligand complexes deposited in the PDB database.

[16] An example of metal interaction is the N–Fe bond in the Heme.

sulfurs (when not hydrogen bond donors or acceptors) and carbons (when are not bound to a heteroatom). The metal interaction term is a simple contact term. One interesting term is the frozen rotatable bonds term that represents the entropy loss of the ligand upon binding. A frozen rotatable bond is a non-terminal bond, in which the atoms on the both sides of the bond are in contact with the protein. The regression for calculating the relative weights of the single terms was calculated on a training set of 82 X-ray complexes (17 aspartic peptidases, 15 serine peptidases, 15 metallo peptidases, 16 sugar-binding proteins and 19 other proteins).

The LIECE scoring function is a hybrid between a force-field and an empirical scoring function, because it consists of the weighted contributions of van der Waals and electrostatics (the sum of coulombic interactions and desolvation penalty) to the binding free energy:

$$\Delta \mathrm{G}_{bind} = \alpha \Delta \mathrm{E}_{vdW} + \beta \Delta \mathrm{G}_{elec}$$

It differs from a pure force-field based scoring function because of the regression based weighting of the force-field energy contributions. Differently from ChemScore, LIECE is not trained on several different targets, but on the protein target only, when two conditions are satisfied: first, one co-cristallized X-ray structure of an inhibitor must be available; second, at least a series of inhibitors (preferably congeneric to the cocrystallized one) of which the binding energy has been experimentally determined has to be available. If other series of inhibitors are available and their binding mode can be determined by docking, they can be employed, too, with some caveats. The regression for calibrating LIECE's weights (α and β parameters) is then done on all the known inhibitors. Ideally, for achieving the best predicting ability, the inhibitors potency should span several orders of magnitude, from very low nano-molar to high micromolar, and the inhibitor molecular weight distribution should be similar to the one of the library that is being scored. An example of a LIECE model I built for BRAF kinase is presented in figure 1.8.

One disadvantage of LIECE is that its predicting power decreases for proteins with a large active site, when the training set for the regression does not explore all sub-pockets. This observation has been made by Dr. Marino Convertino while testing LIECE on his project on β-secretase, whose active site is large and deep. The original training set consisted of large peptidic and peptido-mimetic compounds. The weights determined for large compounds were not particularly suitable for scoring the smaller drug-like compounds he docked. Finally, LIECE is very sensitive to even very small variations of the binding pocket conformation, therefore, one has to check for transferability of the α and β parameters whenever a new structure of

the receptor is employed for docking. While the ChemScore scoring function is transferable among different protein families, LIECE is not transferable, albeit it is still transferable among a single protein family (general LIECE model on kinases[53]). In the general LIECE models developed on three kinases (Cdk2, Lck, and p38), the α and β parameters are α =0.2898±0.0075 and β =0.0442±0.0074. In my LIECE model on BRAF kinase (see figure 1.8), which was not used for the regression of the general LIECE model, the parameters are α =0.2857±0.0120 and β =0.0531±0.0128.

1.5.3 Interpolation vs. extrapolation

The calculated binding free energy of regression-based scoring functions is usually obtained by interpolation, that means that the experimental data cover a certain free energy range in a sparse manner and the predicted energy value is obtained from the regression curve in a portion of space within the aforementioned range.

In the example of figure 1.8, the two inhibitor series (red and black symbols) cover a range of activity from a ΔG of binding of -13 to -5 kcal/mol, but their average ΔG of binding is -10 and -7 kcal/mol. For this reason, the affinity range between -9.5 and -7 kcal/mol, for which just a few compounds are present, will be more likely predicted by extrapolation and not by interpolation, as the figure might wrongly suggest. As for the simpler case of linear regression, the standard error of values derived by extrapolation is higher than for values derived by interpolation.

1.5.4 Knowledge-based scoring functions

Knowledge-based scoring functions are built from statistical analyzes of experimentally determined structures of complexes, assuming that particular frequent interatomic distances, occurring more often than the average, reference value, represent favorable contacts. For the derivations of these functions, one has to define a set of atom types for the protein and for the ligand and has to count all the possible pairs in binned distance intervals. PMF[63] and DrugScore[30] are two well known examples.

In the DrugScore function, potentials were derived for the following atom types: C.3 (carbon sp^3), C.2 (carbon sp^2), C.ar (carbon in aromatic rings), C.cat (carbon in amidinium and guanidinium groups), N.3 (nitrogen sp^3), N.ar (nitrogen in aromatic rings), N.am (nitrogen in amid bonds), N.pl3 (nitrogen in amidinium and guanidinium groups), O.3 (oxygen sp^3), O.2 (oxygen sp^2), O.co2 (oxygen in carboxylate groups), S.3 (sulfur sp^3), P.3 (phosphorus sp^3), F (fluorine), Cl (chlorine), Br (bromine), Met (Ca, Zn,

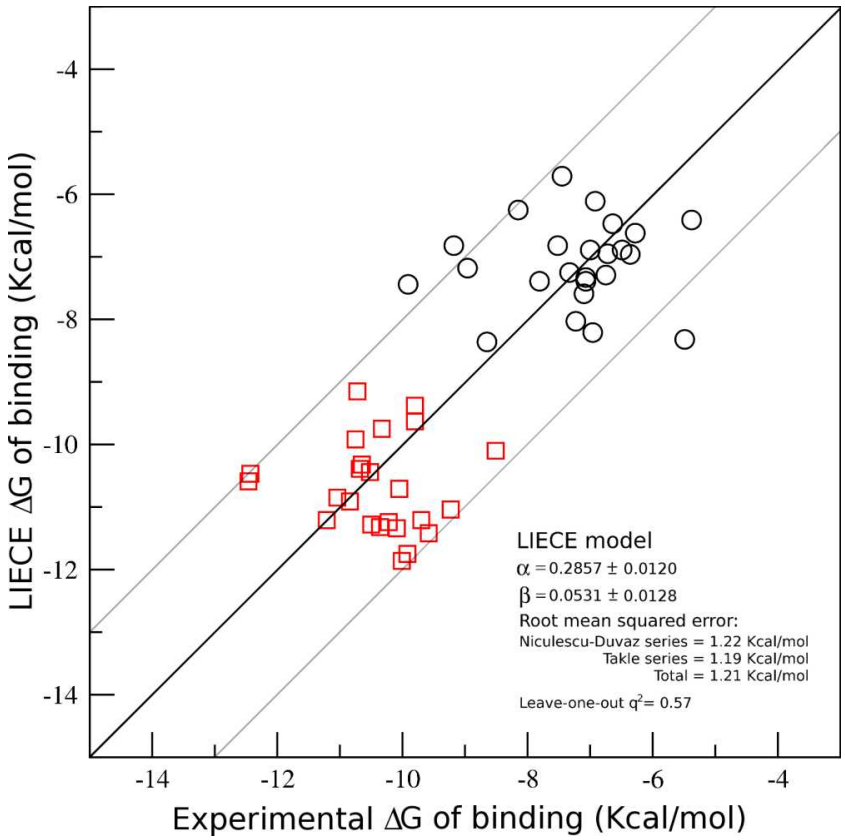


Figure 1.8: A LIECE model of BRAF kinase. The LIECE model has been parametrized using two different sets of inhibitors (red and black symbols in the plot).

Ni, Fe). Some atom types have low occurrence and thus were grouped together: S.2 (sulfur sp^2) and S.3, N.4 (positively charged nitrogen) and N.3. To derive potentials in DrugScore, a non-redundant set of protein-ligand complexes was extracted from ReLiBase[17], an annotated database of receptor-ligand complexes.

In figure 1.9 some statistical preferences are represented. The preferences can be divided into two classes: polar and apolar interactions. The polar ones have a marked maximum for distances between atom pairs of 2.5–3 Å and the apolar are significally preferred with respect to the reference state for distances between atom pairs of 3.5–6 Å. The reference state is calculated as arithmetic mean over all normalized pair correlation functions. It may be regarded as a mean interaction preference between

[17] http://www.ccdc.cam.ac.uk/free_services/relibase_free/

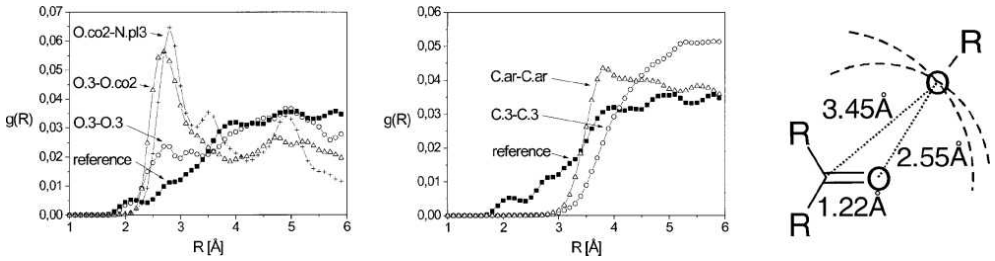


Figure 1.9: DrugScore statistical preferences. Frequency $g(r)$ of polar/charged (left), non-polar and aromatic interactions (center). The optimal geometry of a carbonyl-alcohol hydrogen bond can be deduced from the statistical preferences (right) as a function of their distance. The peak of the C.2-O.3 preference is at 3.45 Å, the one of O.2-O.3 is at 2.55 Å and the equilibrium distance of the C.2-O.2 bond is 1.22 Å. The resulting calculated C.2-O.2-O.3 angle is 128 degrees, in agreement with the values in the PDB database. Therefore, the directionality of the bond is implicitly contained in the statistical preferences. Reproduced from [30].

"averaged atom-types", thus mainly representing non-specific contributions from dense packing.

The advantage of knowledge-based scoring functions is that they are more transferable. Moreover, desolvation, entropic and other effects that drive binding are implicitly taken into account by the procedure. Less populated states are given less statistical preference, therefore probably avoiding docking artifacts.

Finally, while protonation states are important for methods such as LIECE, which are based on rigorously calculated electrostatics, they are not problematic for scoring functions like DrugScore or PMF, because they are not taken into account for building the model as the assignment of atom types does not require knowledge of the protonation state.

1.6 Assessing virtual screening performance

The performance of a virtual screening campaign can be evaluated when the number of active compounds in the database is known or can be reliably estimated. Usually, only a small subset of compounds, such as the 50–100 best ranking compounds, is selected for experimental tests.

Enrichment factors

The Enrichment Factor of a selected subset of a database, usually the 5% or 10% ($EF_{5\%}$ and $EF_{10\%}$, respectively), is the ratio between the percentage of active compounds in the selected subset and the percentage of the entire database:

$$EF_{\%} = \frac{Hits_{selection}}{Hits_{total}} \cdot \frac{NC_{total}}{NC_{selection}}$$

where NC_{total} is the number of compound of the database and $NC_{selection}$ is the number of compounds in the selected subset. $Hits_{total}$ is the whole number of hits in the database and $Hits_{selection}$ is the number of hits in the selected subset.

For example, let's assume that 20000 randomly chosen molecules from a chemical library form a database of not active compounds towards a particular target. Known inhibitors of this target, for example 50, are included into this database and, then, the whole database, consisting of known active and inactive molecules, is screened by docking and scored. A perfect docking and scoring would rank the 50 known inhibitors in the first 50 places. Unfortunately, this is rarely the case and docking and scoring programs must be assessed for their ability of fishing active compounds out. In the first 10% of the database there are 2005 molecules, and only 27 known inhibitors are correctly retrieved. The enrichment factor would be

$$EF_{10\%} = (\mathbf{27}/50)(20050/2005) = 5.4$$

Importantly, the maximum theoretical enrichment factor is the percentage at which it is calculated. In the aforementioned example:

$$EF_{10\%,\,max} = (\mathbf{50}/50)(20050/2005) = 10.0$$

The performance of several docking and scoring algorithms can be therefore assessed and compared for the highest enrichment factor.

Enrichment curves

Enrichment factors do not consider how fast known active molecules are retrieved from a pool of inactive ones and the same enrichment factor can therefore be obtained if the active compounds are grouped together in the best positions of the list, or if they are sparse in the analyzed list, or, in a limit case, if they are all grouped together at the bottom of the analyzed list.

Enrichment curves are useful for coping with this issue. Enrichment curves report the percentage of active compounds as a function of the fraction of screened library. The steeper is the curve and larger is the area under

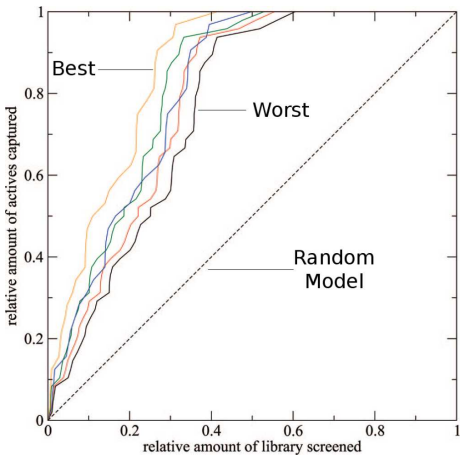


Figure 1.10: Enrichment curves. Five enrichment curves are plotted. The steepness and the area under the curve of the yellow curve is bigger than the one of the black curve, and therefore its model is better at finding true active molecules. Reproduced from [53]

the curve, the better is the procedure at finding true active compounds (see figure 1.10).

1.6.1 Consensus scoring

It appears natural that combining results from such a number of scoring functions can lead to an increase in scoring performance[23]. Intuitively, if several different scoring functions always put a certain compound in the first ranked positions, it is more likely that it is a true active, than a compound that is always scored low. For example, an evaluation of docking results of three popular docking programs with seven scoring functions[3] found an improvement with respect to the docking score of the of circa 10% with the best scoring function, 25–40% with the two best scoring functions, and 65-70% with the three best scoring functions. A successful use of consensus scoring has been reported in a high-throughput docking study that resulted in the identification of inhibitors of plasmepsin[25]. Consensus scoring is therefore a very powerful tool for increasing the success rate rate of docking campaigns.

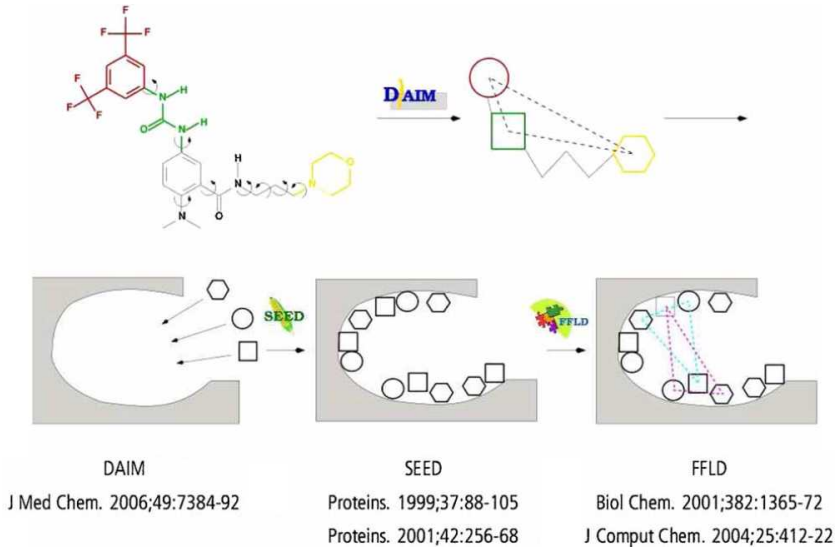


Figure 1.11: The in-house docking procedure. First, the molecule is divided into three fragments by DAIM; second, the fragments are docked into the active site by SEED; third, the molecule is placed into the active site by FFLD, using the precalculated fragment positions as anchors to drive the positioning. Reproduced from [53].

1.7 An example of a docking approach

A docking strategy example will be given based on the docking suite developed in my research group. The underlying idea of this particular approach is the fragment based drug-design by linking ligand fragments. Briefly, if fragments are known to (weakly) bind a protein, linking them will most probably generate a better binder. Following this framework, the in-house docking approach takes advantage of the interplay of three programs:

- DAIM[51] (**D**ecomposition **A**nd **I**dentification of **M**olecules)
- SEED[58] (**S**olvation **E**nergy for **E**xhaustive **D**ocking)
- FFLD[8, 14] (**F**ragment-based **F**lexible **L**igand **D**ocking)

In this docking procedure, a molecule is fragmented and the best positions of the fragments in the binding pocket drive the optimal placement of the whole molecule (see figure 1.11). The real generation of novel molecules by linking fragments is briefly illustrated in the *de-novo* design (section 1.4).

First, given the 3D structure of a molecule, DAIM automatically generates a triplet of fragments of a given molecule by cutting single, non terminal

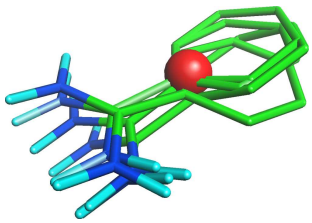


Figure 1.12: Clustering of benzamidine by SEED. Several positions of benzamidine are generated by SEED in the S1 pocket of Thrombin (not shown) and they are clustered together. The poses shown here are the members of the most favorable cluster. The sphere in red is the geometrical center of the cluster.

rotatable bonds and by avoiding to remove functional groups. Valences are then filled either with a hydrogen, or with a methyl, if adding a hydrogen would create a hydrogen bond donor which was not present in the original molecule. Of the all possible fragments that can be obtained by single bond cleavages, DAIM automatically chooses three fragments to represent the pharmacophore of the molecule.

Second, SEED exhaustively docks the three fragments to a binding pocket with an accurate force-field based scoring function comprising desolvation energy. The fragment is docked in all the possible positions in the volume of the binding pocket and that particular attention is given to the hydrogen bonds, for which all the possible reasonable solid angles between the donor and the acceptor (D-H$\cdots$A) are sampled. Fragments and receptor conformations are kept rigid during the docking. To reduce the complexity of the system, all the docked positions of the fragments (normally several thousands) are clustered together, according to 3D similarity (see figure 1.12) and only the geometrical centers of the most energetically favorable positions are kept as anchors for the final step of the docking procedure.

Third, FFLD generates a number conformations (several hundred of thousands) of the molecule with a genetic algorithm[18] and uses the geomet-

[18] A genetic algorithm (GA) is a search technique used in many disciplines for finding solutions of optimization problems. GAs are inspired by evolutionary biology and from it they borrow words such as chromosome, fitness, mutations, selection, recombination, et cetera. A random generated population of molecules is generated and optimized through many iterations of the algorithm. In the case of FFLD, the "chromosome" of every molecule contains informations on the rotatable bonds angles, therefore the GA perturbates only the conformation of the molecule. The fitness function which is optimized by the GA is a force-field based one (see section 1.5).

rical centers of SEED as anchors to drive the docking of the conformers. Since the same fragment can be present in more than one molecule[19], an important speed-up is gained by using the previously obtained results of SEED for more than one molecule processed by FFLD. It's noteworthy that FFLD can only dock compounds that have at least three fragments because only in this way it can unambiguously place a conformation.

Eventually, the geometry of all the docked poses generated by FFLD is optimized by the program CHARMM[7, 6]. This paper by Huang and Caflisch[41] is a good and comprehensive review of all docking campaigns involving DAIM/SEED/FFLD.

1.8 An example of de-novo design

I will briefly illustrate GANDI (Genetic Algorithm based de-Novo Design of Inhibitors) because I took place to the software testing and therefore also suggested modifications and new features, some of which have been successfully implemented by the developer, Dr. Fabian Dey.

The principal idea of GANDI is growing molecules into the binding pocket of a receptor by coupling previously docked fragments with suitable linkers, both chosen from a pool. The fragments are initially docked with SEED, which ensures their best position into the binding pocket and calculation of an accurate binding energy, and then provided to GANDI. In the example of figure 1.13, three fragments are first docked: a furan and a benzene in hydrophobic pockets of the receptor, and a pyrrole in a polar environment and establishing a hydrogen bond with the receptor. Then, GANDI couples the fragments with linkers, in the example an ester and an amide group. GANDI does not modify the torsional angles of the fragments and the linkers, but many different conformations of both can be provided, therefore providing a certain degree of flexibility. The bond directionality is given by the connection vectors, which are the red arrows in the linkers and the black arrows for the fragments. The connection vectors are usually all the heavy-atom→hydrogen vectors, but they can defined by the user. The difference between fragments and linkers is that linkers must have at least two connection vectors. Therefore, the same molecule can be at the same moment docked with SEED as a fragment and provided to GANDI as a linker, if it has two or more connection vectors. In the example, even if only some are shown, benzene has six connection vectors, furan four,

[19]For example, according to [51], the first ten most frequent cyclic fragments in the 2005 version of the ZINC library are: benzene, chlorobenzene, toluene, fluorobenzene, bromobenzene, nitrobenzene, pidirine, furane, thiophene, and 1,3-dichlorobenzene.

and pyrrole five, one of which is discarded because of the involvement in hydrogen bonding. Importantly, DAIM can fragment molecules and therefore provide GANDI with fragments and linkers whose particular annotated vectors represent the heavy-atom→hydrogen vectors which were involved in single bonds. This way, the problem of the synthetic feasibility of the de-novo generated molecules can be mitigated. Additionally, GANDI contains a list of "prohibited bonds" such as O-O, N-N, N-O, which are not chemically stable. The search for the best combination of linkers and fragments is optimized with a complex, but highly customizable genetic algorithm. The fitness function of each generated molecule for the genetic algorithm is a linear combination of three terms:

$$\mathrm{S}_{total} = w_{ff}\mathrm{E}_{ff} - w_{3D}\mathrm{Sim}_{3D} - w_{2D}\mathrm{Sim}_{2D}$$

where S_{total} is the total fitness and the weights of the terms are expressed by the w. E_{ff} is the force-field calculated energy (see section 1.5), which is the sum of the SEED interaction energy for the fragment and the CHARMM non-bonded interaction energy for the linker. Sim_{3D} and Sim_{2D} are two measures of similarity with respect to some important molecule, such as a known binder, which have been introduced to drive the search towards known molecular scaffold and pharmacophores. The minus signs for the similarity terms are necessary because, during the minimization of S_{total}, E_{ff} becomes more negative and the similarity becomes more positive and closer to the unity (perfect similarity). According to the original paper, the best results have been obtained with w_{ff}=0.02, w_{3D}=1, and w_{2D}=0 by trial and error.

Last, GANDI can be used also in ligand (or lead) optimization, when a particular fragment is known to bind, it is possible to instruct the program to generate molecules that always contain this fragment. In this way, many congeneric compounds are produced and the most interesting ones can be synthesized and tested for activity.

1.9 Molecular Dynamics simulations

The first molecular dynamics simulation of a macromolecule of biological interest dates back to 1977[60] and it started a revolution in thinking of macromolecules, replacing the view of proteins as rigid structures. Nevertheless, already in 1963, Richard Feynman wrote in the *Feynman's Lectures on Physics* the now over-quoted: "Certainly no subject or field is making more progress on so many fronts at the present moment than biology, and if we were to name the most powerful assumption of all, which leads one on

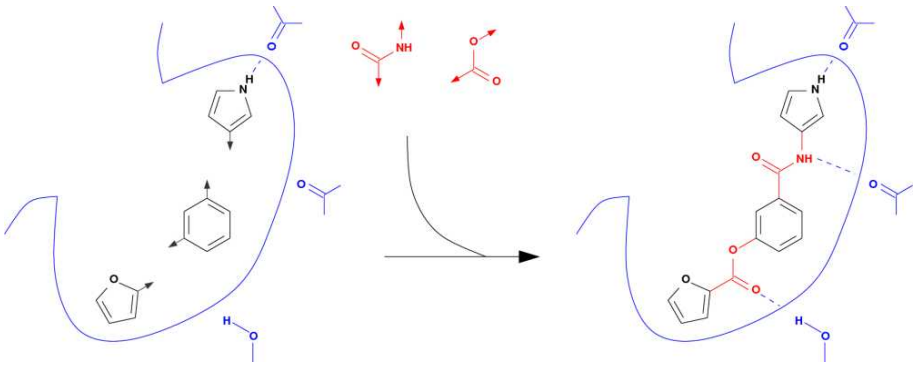


Figure 1.13: Framework of GANDI. Predocked fragments from SEED are linked by linkers from a library. While the position of the fragments is fixed, the linkers are optimized. Reproduced from [19].

and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms". Now, molecular dynamics simulations is a very appropriate tool for studying the *jigglings and wigglings of atoms*.

In structure-based drug design, molecular dynamics simulations can be very useful for elucidating the relationship between structure and function of a protein and for providing alternative conformations of the receptor for docking.

A very interesting review of the role of molecular dynamics simulations in biology is by Karplus and McCammon[48] and a detailed review of molecular dynamics techniques is by Van Gunsteren[37] and Nilsson[65].

CHARMM[7, 6] was used in my research for molecular dynamics simulations performed by classical mechanics, that is they involve the numerical integration of Newton's equations of motion for all the atoms a suitably prepared system. The total energy of the system is empirically expressed as the sum of bonded and non-bonded interactions as a force field in the form

$$\mathrm{E} = \underbrace{\mathrm{E}_{bonds} + \mathrm{E}_{angles} + \mathrm{E}_{dihedrals} + \mathrm{E}_{impropers}}_{\text{bonded interactions}} + \underbrace{\mathrm{E}_{vdW} + \mathrm{E}_{coulomb}}_{\text{non-bonded interactions}}$$

and depends on the bond lengths (E_{bonds}), the bond angles (E_{angles}), the proper and improper[20] torsion (dihedral) angles angles ($\mathrm{E}_{dihedrals}$ and

[20]The improper torsion angle term was introduced to maintain chirality about a tetrahedral extended heavy atom and to maintain planarity about certain planar atoms, such

$E_{impropers}$, respectively), the van der Waals and the coulombic energy (see section 1.5.1 for their analytical expression).

Molecular dynamics simulations can be performed in an explicit or implicit solvent. In an explicit solvent simulation, the protein is immersed in a appropriate water filled box, while, in an implicit solvent simulation, the effects of the solvent are averaged and present as an additional term in the empirical energy function. The water model used by CHARMM is called TIP3P[47], and consists of three interaction sites corresponding to the three atoms of the water molecule. Explicit solvent simulations are more accurate than implicit solvent ones, but also very demanding in terms of computational power, because it is not so uncommon that the protein atoms represent less than 10% of the total atoms of the simulated systems and the time required for the simulation is roughly proportional to the square of the number of atoms.

With the term "minimization" or "energy minimization" it is meant a progressive modification of the coordinates of the system, so that the total energy at the end of the procedure is minimized (i.e. reduced, more favorable). During a minimization, some atomic coordinates can be fixed or biased (by applying a restraint) to a certain position. For example, fixing protein atoms is routinely used in docking, while minimizing a docked pose in the rigid binding pocket, or while optimizing the positions of hydrogens only, for the preparation of a suitable receptor structure.

1.9.1 Preparation of a protein for molecular dynamics simulations

The preparation of a protein for molecular dynamics simulations is not a completely trivial task. Besides simple technical problems arising from the data manipulations of structures obtained from the Protein Data Bank, such as different naming and numbering conventions of CHARMM, many steps require the active participation of the user. For example, crystal structures might have an "alternate location" for some residues, meaning that in the crystal lattice a particular residue assumed two or more conformations. It's then up to the user to decide which position is better, usually the most populated one. If the population of the two states is the same and this residue is supposed to be important for the outcome of the simulation, one possibility would be starting simulations from two protein structures,

as carbons in benzene. Extended heavy atoms are a reduced representation of a molecule, in which non-polar hydrogens are included in the carbon atom. For example, the methylene group, $-CH_2-$, is reduced to a single carbon atom with higher mass and van der Waals radius. With this approximation, the complexity of the system decreases.

each one containing a different rotamer. Additionally, as explained in section 1.3, the crystal structure might contain errors in the positioning of the asparagine and glutamine side-chains, and of the imidazole ring of histidine. Those errors can be found and corrected automatically, but the intervention of the user is always recommended. Moreover, the protonation state of many titrable residues has to be defined, especially if a particular residue is important for the simulated system, and disulphide bridges have to be specified. Hydrogen atoms have to be added, and, while non polar hydrogens are easier to insert (at the equilibrium bond and dihedral angle, and at the equilibrium distance), polar hydrogens are more complicated. In fact, hydrogens belonging to waters or hydroxyl groups can assume different conformations according to the environment. CHARMM copes with this problem by optimizing their initial position with an iterative method which searches for the conformation with the most favorable energy (*hbuild* command).

If some residues are missing in the crystal structure, be they the N- or the C-termini or a flexible loop whose electron density is not observed, to prevent the presence of unwanted, unrealistic charges at the N- or C-termini where the polypeptide chain is interrupted, two solutions are commonly accepted. First, the termini at the missing parts of the protein can be patched with capping groups, such as an acetyl group at the N-terminal part or N-methyl-aminyl group at the C-terminal part, to remove the positive and the negative charge, respectively (see figure 1.14). This is the simplest solution, and it is generally well accepted for the N- and C-termini. Second, the missing segments can be re-introduced by homology modelling. This solution is particularly suitable in case of a missing loop, whose absence can generate an unrealistic system. It is advisable to check the movability of the introduced loop during the simulation, assuming that no electron density of it was determined because of its high flexibility. In fact, if the homology model presents an element of secondary structure in the generated loop which is not present in the "real" structure in solution, the flexibility of the loop can be lower than the expected one. Therefore, the behaviour of the simulated protein structure could strongly differ from the "real" supposed one in solution.

Once the protein structure is prepared, the system has to be minimized (usually in an implicit solvent) until an energy minimum is reached, that is when the sum of all forces acting on every atom is zero. This is due to the fact that in a crystal structure it is very plausible that some atoms have been placed with bond lengths which are a little stretched with respect to the equilibrium value of the force field or that some atoms slightly overlap. If no minimization were carried out and a simulation were started from

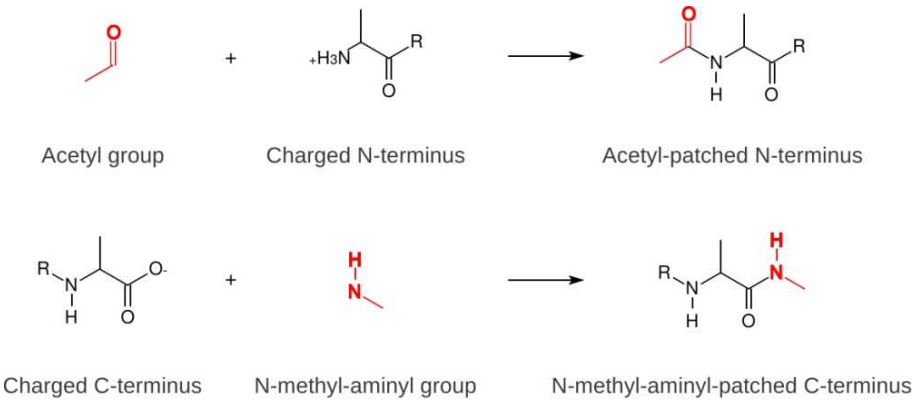


Figure 1.14: Common patches for N- and C-termini. The acetyl group and the N-methyl-aminyl group are two common patches for removing the charge from the N- and the C-terminus, respectively.

a structure whose energy is distant from a minimum, the aforementioned atoms would be subject to large forces and they would accelerate to velocities that could not be handled by CHARMM. To prevent big movements of the protein backbone upon minimization, gentle constraints are placed upon backbone atoms, so that their position will not be modified too much, but allowing an energy minimum to be reached.

Crystallization waters and ions which appear in the crystal structure should be retained if they are buried[68], or if they are known to be biochemically important. For example, in figure 1.15 are shown the effects of the removal of buried water. After minimization, two arginine residues originally making hydrogen bonds with cocrystallized water (magenta) change their position (blue), filling the void left by missing water molecules. Consequently, the algorithm which inserts the protein in a water box will not be able to place water in the previously occupied positions.

1.9.2 Setting up an explicit solvent molecular dynamics simulation

In explicit solvent molecular dynamics simulations, after the preparation of the receptor structure, the protein is immersed in a box of water. Salts (NaCl or KCl) are added to the system for neutralizing the net charge of the protein and for reaching the the physiological ionic strength (0.2M). In fact, long range electrostatics effects might influence the stability of the protein if ions are not present[43].

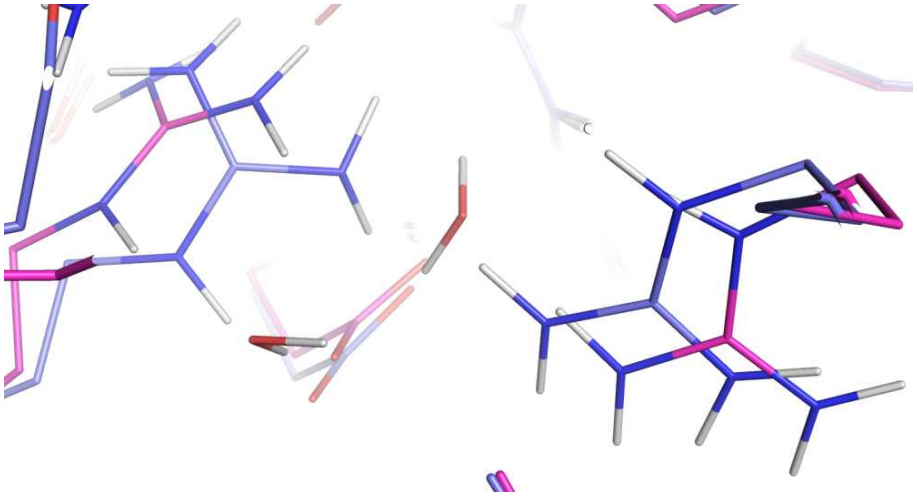


Figure 1.15: The effect of the removal of buried water. After minimization, two arginine residues originally making hydrogen bonds with cocrystallized water (magenta) change their position (blue), filling the void left by missing water molecules.

To prevent finite size problems due to the interface between the box of water and the vacuum, periodic boundary conditions are employed, which eliminate interface problems by surrounding the simulated systems by images of itself. The easiest way to accomplish this task is to have cubical symmetry[21] and surrounding a cubic water box with other 26 images in the x, y, and z directions (see figure 1.16 for a 2D example). According to this method, boundaries are no more present and atoms close to the borders of the box interact with others from the image but the number of interactions that have to be calculated increases. To prevent that the protein interacts with itself from an image (thus simulating a crystal), the water box size has to be big enough. Usually, the size of the box is so that a minimum distance of 13–16 Å is left from every atom of the protein to the borders.

After a fully solvated system is prepared, the position of water molecules is deeply minimized, usually fixing the heavy atoms of the protein in place, to optimize the hydrogen bonding network.

Implicit solvent molecular dynamics simulations do not need any solvent or counter ion, just the choice and the setup of the most suitable implicit solvent. For example some implicit solvents are more appropriate for small peptides and some others implement routines for the emulation of mem-

[21] Other symmetries are supported by CHARMM such as rhomobohedra, rhombic dodecahedra, truncated octahedra, and hexagonal prisms.

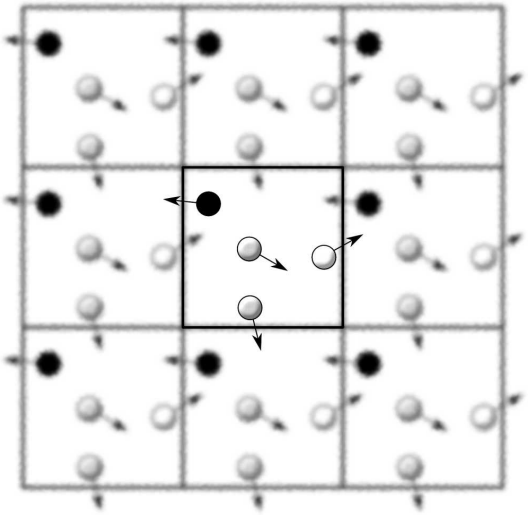


Figure 1.16: Periodic boundary conditions. To prevent finite size problems of a a protein immersed in a box of water surrounded by empty space, periodic boundary conditions are employed. The atoms in the central box interact with identical images of themselves along all directions. Moreover, if a particle exits the simulation box from a side, it reenters from the opposite side, as shown in the figure by arrows.

branes. Simulations in implicit solvents are faster but less accurate than simulations in explicit solvents.

1.9.3 Running a molecular dynamics simulation

A simplified flowchart of a molecular dynamics simulation is presented in figure 1.17. According to it, a molecular dynamics simulation can be divided in different phases. First, the setup of the system, the choice of an integration step Δt (explained later), and assignment of the initial random velocities. Importantly, the two systems evolve differently, if two simulations are started with different initial velocities. Second, the calculation of the forces acting on the atoms and therefore the accelerations of the atoms, by numerical integration of the Newton's law in the differential form. The simulation time step is the length of the integration step. For each integration step, at least one energy calculation is required. Third, the movement of the atoms according to the calculated accelerations and the chosen time step. Fourth, the simulation time is increased. Then the point 2–4 are iterated many times until a particular simulation length is reached.

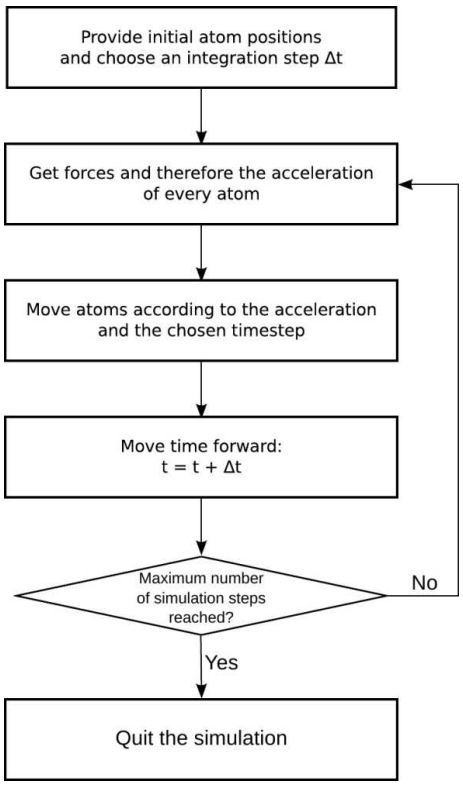


Figure 1.17: A simplified flowchart of a MD simulation.

A very important parameter is therefore the integration time step. For sampling the system correctly[22], the time step should be smaller than the highest vibration frequency (that is the smallest period) of the system, which is the C-H bond stretch for biopolymers. The problem is that the simulation time is practically directly proportional to the integration step: if one needs to collect 1 ps simulation time and the integration step is 0.5 or 2 fs, he has to run the simulation for 2000 or 500 steps, respectively. One very common approximation for increasing the integration step is to keep the bond length of hydrogen atoms fixed, therefore allowing integration steps of 1 or 2 fs.

[22]Just imagine that you want to sample the position of the sun in the sky. You can determine it too frequently, such as every second, or too rarely, such as every day. In the first case, the system is ideally sampled correctly, because the observation frequency is much higher then the frequency of the event. In the second case, the system is not correctly sampled, because its frequency is much higher then the observation frequency, and therefore the sun would seem (almost) not to move in the sky.

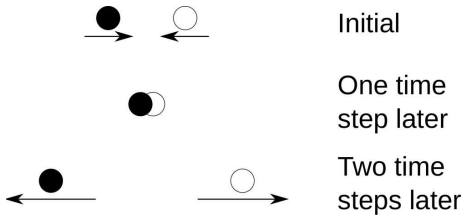


Figure 1.18: The effect of a too large integration time step. If the integration time step is too large, atoms can interpenetrate their van der Waals radius generating high potential energy, which is transformed into kinetic energy, provoking a violent repulsion.

A problem that can arise if the integration step is too large is the kinetic energy instability of the simulation (see figure 1.18). If the time step is too large, two atoms can be displaced too much and therefore they can interpenetrate (third step of figure 1.17). As a consequence, a very high potential energy will be generated, which will result in a high acceleration and thus high kinetic energy. The accelerated atoms will very probably clash with some other atoms and the system will be soon unstable[23].

Heating, equilibration and production phases

A molecular dynamics simulations consists of three phases.

During the *heating phase*, the temperature of the system is gradually increased from 0 K to the desired one for the production phase, usually 300 or 310 K. The temperature of a system made up by N atoms is related to the kinetic energy, calculated from the atomic velocities, by the equation:

$$\frac{3}{2}\mathrm{N}k\mathrm{T} = \sum_{i=1}^{\mathrm{N}} \frac{m_i v_i^2}{2}$$

The initial random velocities are randomly assigned according to a gaussian distribution whose mean is the desired temperature. During heating, temperature rises by progressive reassignments of the velocities and the whole heating process lasts 0.25–0.5 ns. Usually, the protein atoms are either kept fixed during heating or a constrain on their position is applied.

The *equilibration phase* starts when the desired temperature is reached. During equilibration, velocities are frequently rescaled (not reassigned) to maintain the average temperature close to the wanted one. The constraints

[23] When a system becomes unstable, it is informally said that it *explodes*.

on the protein atoms are first gently released and then completely removed for a certain length of the equilibration phase. The equilibration process lasts 0.5–1.0 ns.

During the *production phase* all the constrains on the atom positions are removed and the system evolves in absence of velocity adjustments.

To run a molecular dynamics simulation, whose result can be compared to experiments, a NpT (isothermal-isobaric) ensemble should be generated. In this ensemble, the number of particles of the system is fixed, and the pressure and the temperature are maintained by a barostat and a thermostat, respectively.

The duration of the production phase depends on the observed phenomenon. For example, the analysis of local motions such as atomic fluctuations, side chain motions, and loop movements requires less simulation time than the analysis of large scale motions such as folding/unfolding and dissociation/association of proteins. The necessary simulation length can not always be determined in advance, unless, for example, the time scale of a phenomenon (nanoseconds, microseconds, milliseconds...) is known from experiments.

The positions of the atoms of the system, recorded at regular intervals (usually 1–2 ps), generated a trajectory.

1.9.4 Properties easily calculated from molecular dynamics simulations

From the trajectory of a molecular dynamics simulation, several properties can be easily calculated.

Average energy

The average energy of a trajectory of N frames is

$$\langle \mathrm{E} \rangle = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathrm{E}_i$$

The average energy can be useful to compare different simulations.

In the LIE method for the calculation of the binding energy of a ligand to the protein (see section 1.5), the interaction energy of the ligand is average calculated along the molecular dynamics trajectory.

RMSD between two structures

The Root Mean Squared Distance between two structures such as the X-ray structure and a simulation frame, or a docked pose and a native pose is

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}(x_i - x_{i,ref})^2 + (y_i - y_{i,ref})^2 + (z_i - z_{i,ref})^2}$$

where N is the number of atoms; x_i, y_i, z_i are the coordinates of the atom i after best superposition on a reference structure; $x_{i,ref}$, $y_{i,ref}$, $z_{i,ref}$ are the coordinates of the atom i in the reference structure.

RMSD expresses how different is an object with respect to another after the best superposition of the two. A RMSD value of zero means perfect superposition. It can be used for inferring the stability of a simulation if RMSD between the simulation and the minimized structure which entered in the dynamics is usually lower than 3–4 Å.

RMSF and B-Factor

The Root Mean Squared Fluctuation of an atom i in a trajectory segment of N_f frames is

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{\mathrm{N}_f}\sum_{i=1}^{\mathrm{N}_f}(x_i - \tilde{x})^2 + (y_i - \tilde{y})^2 + (z_i - \tilde{z})^2}$$

The coordinates $\tilde{x}$, $\tilde{y}$, and $\tilde{z}$ refer to the average structure calculated along small trajectory segments (1 or 2 ns) after superposition.

RMSF is a measure of local atomic flexibility (due to the averaging along short trajectory segments) and it can be related to the crystallographic B-factor by

$$\mathrm{B}_i = \frac{8\pi^2}{3}(\mathrm{RMSF}_i)^2$$

An example of RMSD and RMSF calculations is shown in figure 1.19.

Finally, often, it could be informative to calculate the residence time of water molecules in the active site, because their position could be a *hot-spot* and therefore their position might retained for docking purposes[38].

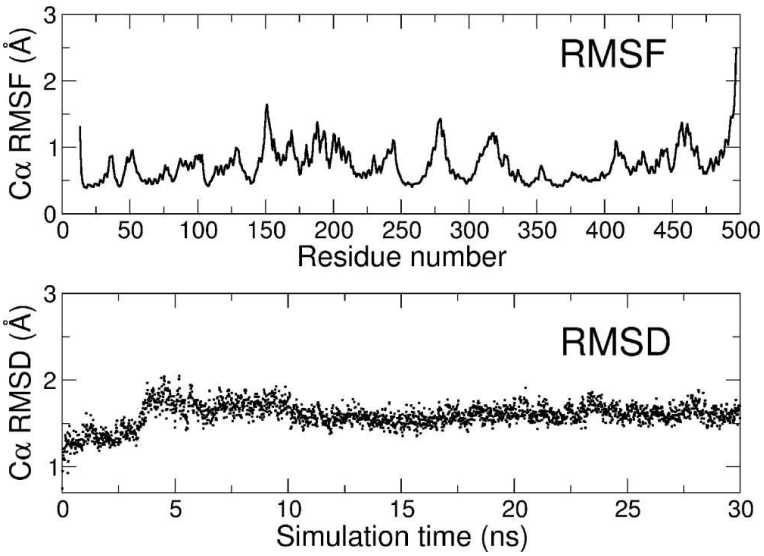


Figure 1.19: RMSF as a function of residue number and RMSD as a function of simulation time. C_α atoms were used and a trajectory of 30 ns with averaging every 2 ns for the RMSF. High values of RMSF indicate structural flexibility for the corresponding amino acids and high values of RMSD for a frame indicate structural distance from the template of the conformation of the protein in this frame.

1.9.5 An application of molecular dynamics simulation for virtual screening

An interesting paper by Ekonomiuk and Caflisch[21] presents an application of molecular dynamics simulation for the selection of an appropriate new conformation of an enzyme for docking. They ran a 1 ns molecular dynamics simulation of the West Nile virus nonstructural 3 protease and extracted 100 equispaced snapshots.

To choose the most appropriate non-native snapshot for docking, they docked to every snapshot three small fragments with SEED and chose the snapshot in which the binding energy of all three fragments was most favorable (see figure 1.21). In a sense, it was not the protein to select the binder, but the binders (although not known for binding) to select the protein. The three fragments were benzene, methylguanidinium, and 2-phenylimidazoline (see figure 1.20).

Benzene was chosen because it is the most frequent fragment in known drugs. Methylguanidinium and 2-phenylimidazoline were employed because the active site of the protein displays many negatively charged residues and

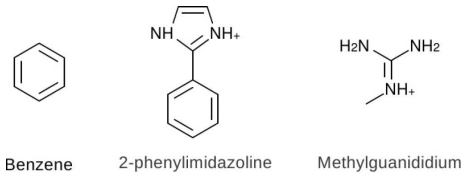


Figure 1.20: Fragment for the choice of the snapshot. Benzene, methylguanidinium, and 2-phenylimidazoline were docked with SEED. Benzene is the most common fragment in known drugs; methylguanidinium, and 2-phenylimidazoline were employed because the active site of the protein displays many negatively charged residues and hydrogen bond donors.

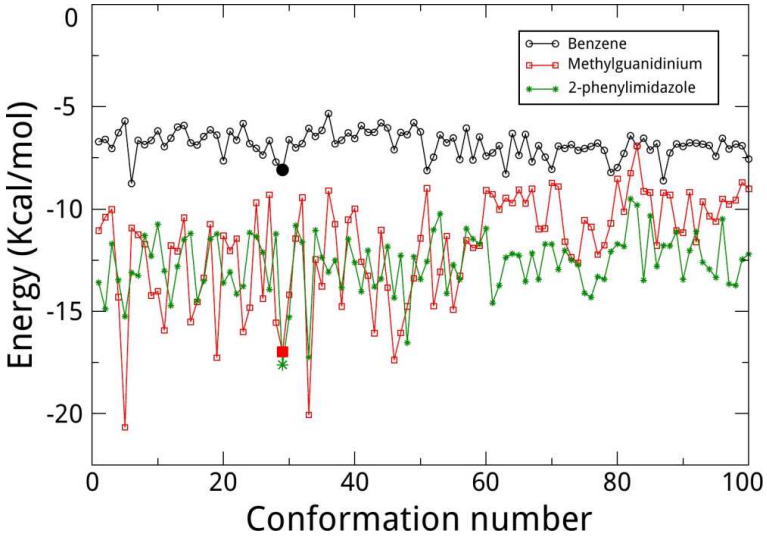


Figure 1.21: Choice of a snapshot from a simulation. A 1 ns molecular dynamics was run and 100 snapshot extracted. The most suitable snapshot for docking is the one in which the binding energy of three fragments (benzene, methylguanidinium, and 2-phenylimidazoline) is the most favorable (snapshot 29). The horizontal lines represent the binding energy in the X-ray structure. Reproduced from [21].

hydrogen bond donors. Of course, fragments can come from other sources, such as natural ligands, known inhibitors, and preferred pharmacophores of the protein. Importantly, the first two sources can introduce a bias for the selection of the snapshot.

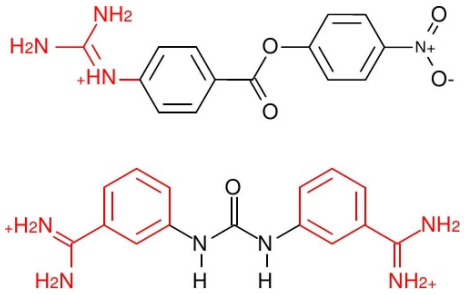


Figure 1.22: The two inhibitors of West Nile virus. The methylguanidinium and the benzamidine fragments are shown in red.

Interestingly, the two inhibitors they found using this strategy would not bind to the native X-ray conformation, confirming again the importance of induced fit or conformational selection (see section 1.3.5) problems in docking.

The choice of methylguanidinium and 2-phenylimidazoline, for selecting the most appropriate protein conformation, was also reasonable, because one of the two inhibitors identified in [21] contains methylguanidinium and the other benzamidine, the latter being structurally very close to 2-phenylimidazoline (see figure 1.22). It is remarkable that the docked library was not biased towards any of the fragments used for selecting the snapshots, but, probably, the protein conformation was "biased" towards the three aforementioned fragments.

In a follow-up paper[20] on longer molecular dynamics simulations of the same protein used for docking, Ekonomiuk and Caflisch advanced the hypothesis, supported by computational evidence, that conformational selection and not induced fit drives the binding of ligands to the West Nile virus protease.

Chapter 2

Allosteric modifiers of Cathepsin K

This work, a collaboration with Dr. Marco Novinec, PostDoc in the group of Prof. Antonio Baici at the UZH, represents a recent (and still in early stages) example of virtual screening applied to the search for allosteric effectors of a peptidase. The main challenge of this project is the absence of a known allosteric binding pocket of cathepsin K, and therefore the reasonable definition of a putative one.

2.1 Allostery

Allostery can be defined as the change in ligand affinity or catalytic efficiency of a protein triggered by the binding of a ligand to a site distant from the active site, often called *exosite*. The current view of allostery is that proteins exist as a population of many dynamically linked states and that the ligand binding shifts the equilibrium towards a particular state. Three good reviews about protein allostery are [50, 35, 73].

Allostery is very interesting for drug design. One benefit of targeting an allosteric site would be specificity. For example, inhibitors of some protein families, such as kinases, are not very specific because of the high conservation of the active site pocket which binds ATP. Another benefit would be that an allosteric inhibitor would very probably behave as a hyperbolic inhibitor. Hyperbolic inhibitors reduce the enzymatic activity without abolishing it completely at saturating conditions and therefore they have a very important pharmacological potential of maintaining the level of enzyme which is necessary for the normal physiological activity of the organism.

2.2 Cathepsin K

Cathepsin K is a cysteine peptidase of the papain family and it is interesting from therapeutic point of view because it is involved in bone turnover. A mature cathepsin K consists of a single polypeptide chain of 215 residues. Inhibitors of cathepsin K have been developed for the treatment of osteoporosis. A lot of effort has been spent at designing specific and efficient inhibitors of cysteine peptidases in general, and most of them target the active site of the enzymes.

In a very recent paper, Marko Novinec and colleagues showed that cathepsin K fluctuates between multiple conformations that are differently susceptible to macromolecular inhibitors and can be manipulated by varying the ionic strength of the medium[66]. Moreover, they found that glycosaminoglycans, such as chondroitin sulfate and dermatan sulfate, which are long unbranched negatively charged polysaccharides, are natural allosteric modifiers of cathepsin K in enzymatic assays performed with low-molecular weight substrates such as short peptides.

2.3 Binding pocket definition

The goal of the project is to find an allosteric modifier of cathepsin K, therefore the active site of the enzyme can not be targeted. The binding pocket definition was carried out first by the statistical coupling analysis (SCA) method and then independently confirmed by normal mode analysis and principal component analysis.

2.3.1 SCA method

The putative allosteric binding pocket of cathepsin K was determined by Dr. Marko Novinec analyzing the family of papain-like cysteine peptidases with the SCA method, developed by Ranganathan and coworkers[57, 72, 36]. It consists of the analysis of the sequence conservation of amino acids among a protein family. Instead of analyzing the evolutionary conservation of a single amino acid, they study pair-wise and even higher order coupling of conserved amino acids, because protein structure and function also depend on the cooperative action of amino acids. Indeed, such analyzes of correlation have contributed to the confirmation and the identification of allosteric mechanisms in proteins.

For example, Ranganathan and coworkers found a conserved networks of residues connecting the binding site of a PDZ domain to a distant patch on

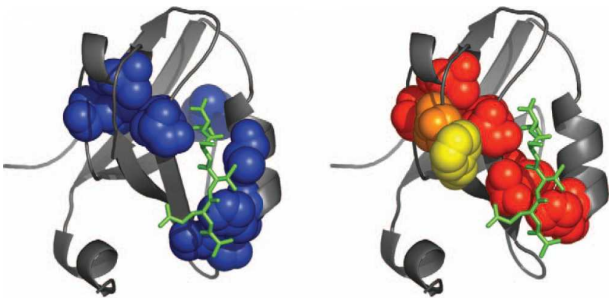


Figure 2.1: Network of conserved residues in a PDZ domain. The network of conserved residues connecting the binding site of the PDZ domain to a distant patch found by SCA (left) and NMR (right). Reproduced from [73].

the surface of the protein[57]. The same patch and the network of residues was then found by a NMR comparison of the side-chains dynamics of the same PDZ domain in the bound and unbound states[28]. The binding of the polypeptide to the PDZ domain has an effect at a site which is distant from the active site. Figure 2.1 shows the network of interacting residues as found by SCA and NMR.

The principal outcome of the SCA performed by Dr. Marko Novinec was a pathway of aminoacids connecting the active site of cathepsin K to two distal sites on the protein surface (see figure 2.2). One site corresponds to a small pocket and therefore was targeted for virtual screening studies, hypothesizing that an interaction of a molecule with this site might have a consequence on the active site.

The other site corresponds to the binding pocket of glycosaminoglycans, natural allosteric effectors of cathepsin K. A structure of cathepsin K in complex with chondroitin sulfate is available (PDB accession code: 3H7D) but the enzyme surface where the polysaccharide binds was not used for docking because it is shallow and its binding is driven by many salt bridges between negatively charged sulfate and carboxy groups of disaccharide units and lysine and arginine residues of cathepsin K (see figure 2.3). Interestingly, this binding surface was also identified by SCA before the structure was released.

2.3.2 Normal mode analysis

The vibrations of a molecule are determined by its *normal modes* and they are calculated by normal mode analysis (NMA). As an example, the vibra-

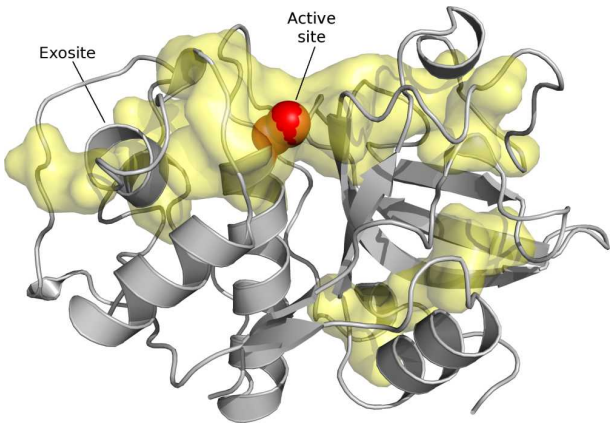


Figure 2.2: Network of conserved residues of cathepsin K found by SCA. The network of conserved residues is in yellow, the cysteine of active site of the enzyme is in red.

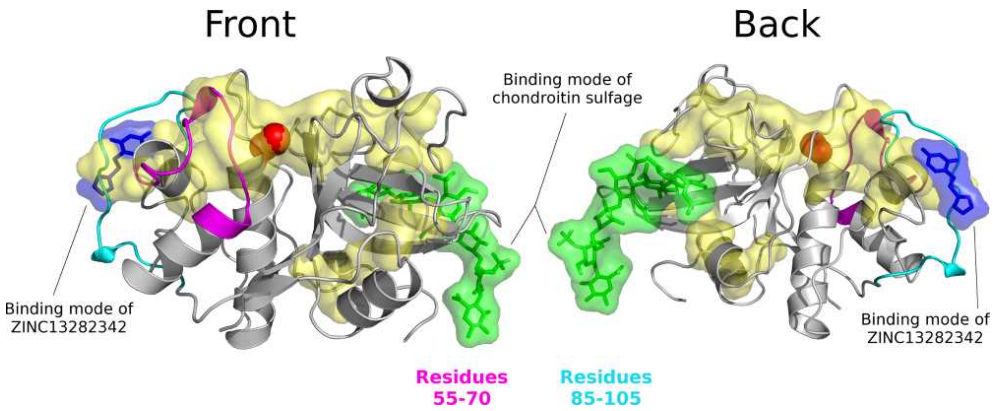


Figure 2.3: Binding mode of glycosaminoglycans. The representation from the back is obtained by a 180 degrees rotation around the axis of the α-helix which contains the cysteine of the active site. Chondroitin sulfate is in green and one allosteric effector found by docking is in blue. The two polypeptide segments indicated in figure 2.7, namely residues 55-70 and 85-105, are in magenta and cyan, respectively.

tions of a small linear molecule such as carbon dioxide are described by four normal modes: the symmetric stretch, the asymmetric stretch, and the two bends (see figure 2.4). The two bending modes, which differ only by their direction, have the same energy. In general linear molecules have 3N-5 normal modes, where N is the number of atoms, and non-linear molecules have

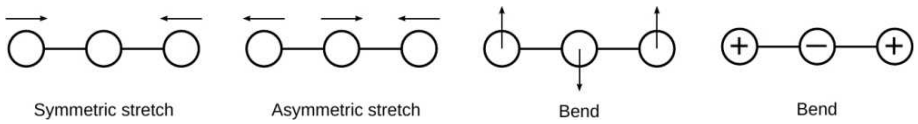


Figure 2.4: Normal modes of a linear molecule.

3N-6 normal modes. The normal modes of simple molecules can be used for the interpretation of IR spectra because the simplest IR bands appear from the normal modes of molecules.

The -5 and -6 subtractive terms for the calculation of the normal modes are due to the fact that 5 modes for linear molecules and 6 modes for non-linear molecules are *trivial*, that is they are pure only rotational or translational movements of the whole molecule and therefore they do change the energy of the molecule. The classical treatment of molecular vibration approximates each mode as a simple harmonic oscillator, that is that the oscillation is subject to the Hooke's law ($F=-kx$). It is possible to extend the calculation of the normal modes to big macromolecules such as biopolymers, but it is necessary to employ an implicit solvent. The requirement of the implicit solvent might therefore introduce a further approximation. Moreover, the method is strongly depending on the force-field and the implicit solvent, if any, used for the calculations and the results may vary for different setups. Nevertheless, NMA is very useful for biological systems and a very comprehensive review on the usage of normal modes analysis for biomolecular systems is from Bahar and colleagues[2].

Importantly, low frequency normal modes are linked to long time scale movements of macromolecules, such as loop motions and domain motions, and therefore conformational changes involved in allostery can be predicted by normal modes analysis because many allosteric proteins are constructed from several subunits linked by hinges that can move relative to each other. For example, the cumulative sum of the first 15 modes of Myosin V can describe the 71% of the rigor to post-rigor transition in either directions[13].

The binding pocket found by the SCA method was also involved in the first (lowest frequency) normal modes (see figure 2.5). To obtain a structure which is perturbed according to the direction of the normal modes, normal modes have to be "projected". In the projection of the two first normal modes there is an almost rigid rotation of the two lobes of the protein (right and left lobe with respect to the α-helix where the cysteine of the active site lies), so that the helix indicated by A in figure 2.5 occludes the active site. The occlusion of the active site might have effects on sub-

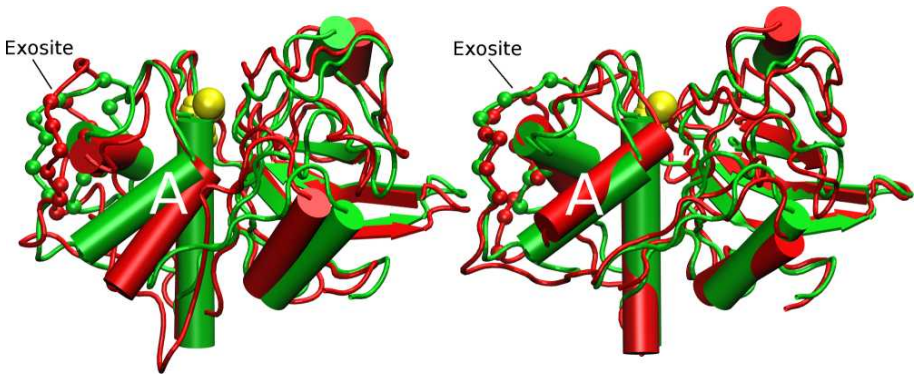


Figure 2.5: Projection of the first two normal modes of cathepsin K. The starting X-ray structure is in green and the structure obtained by the projection of the first normal mode is in red. The active site cysteine is in yellow CPK representation. In the first two normal modes, the helix A occludes the active site of the protein. Green and red beads show the Cα atoms of the binding pocket residues.

strate binding and therefore on the activity of the enzyme. Interestingly, the movement of the helix A is coupled with the deformation of a polypeptide loop which is part of one of the two binding pockets identified by SCA. Assuming that the harmonic approximation in the NMA holds, it can be hypothesized that the interaction of a small drug-like molecule with the protein in the surface of the exosite might influence the movement of the aforementioned helix and therefore could have an effect on the activity of the enzyme.

2.3.3 Principal component analysis

Principal component analysis (PCA), also called essential dynamics, is a mathematical procedure that allows to separate "essential" motions from "non-essential" ones of a molecular dynamics trajectory and therefore can identify collective motions of the trajectory. It is based on the diagonalization of the covariance matrix of the atomic fluctuations during a molecular dynamics simulation. The covariance of the atomic fluctuation is calculated with respect to an average structure obtained from the trajectory. A set of eigenvalues and eigenvectors are generated by the the diagonalization procedure. The eigenvectors describe the concerted motions of atoms in the Cartesian space, and the eigenvalues are the amplitude of the motion.

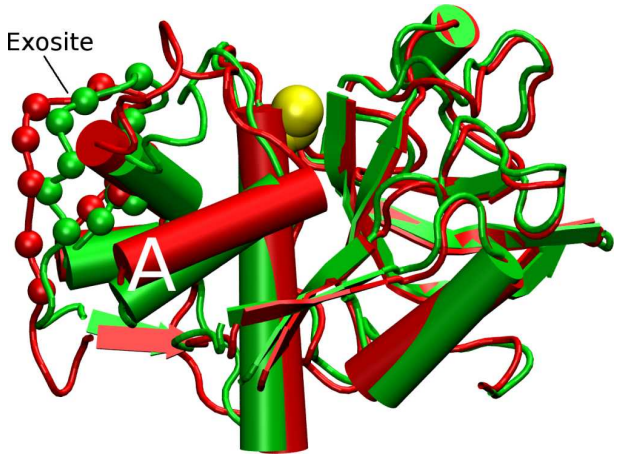


Figure 2.6: Projection of the first eigenvector of the PCA on cathepsin K. The starting X-ray structure is in green and the structure obtained by the projection of the first eigenvector is in red. The active site cysteine is in yellow CPK representation. In the projection, the helix A occludes the active site of the protein. Green and red beads show the Cα atoms of the binding pocket residues.

The main hypothesis of PCA is that the motions of the atoms along eigenvectors with large eigenvalues are meaningful for describing the significant motions of a protein. A new trajectory which describes the motion along an eigenvector can be obtained by projecting the molecular dynamics trajectory onto that eigenvector. Contrary to normal mode analysis, the PCA method does require that the motions of the atoms are completely harmonic because it is performed on MD trajectories. Second, the PCA method is also "model agnostic", because it does not require any other information than the position of the atoms in the trajectory. However, a trajectory has to be obtained from a system which is modeled with a particular force-field. Explicit solvent MD simulations, which are more accurate than implicit solvent ones, can be analyzed by PCA, but not by NMA.

PCA was applied on a 300 K 114-ns trajectory of an explicit water MD simulation (data not shown) of the apo-structure of cathepsin K (started from PDB structure 1ATK). The projection of the first eigenvector is very similar to the outcome of the projection of the first two normal modes (see figure 2.6).

2.3.4 Comparison of the three methods

Three very different analysis methods identified the same surface of the enzyme. This finding is remarkable, considering that very different approximations are used: no structural information used in SCA, harmonic approximation in NMA, and a possibly inadequate simulation sampling for PCA (if the simulation time is too short for studying a particular phenomenon). Importantly, the effect of the binding of a molecule to the exosite found by the SCA method can not be predicted. Moreover, kinetic assays can identify a binder only if the compound not only binds to the protein, but also changes the reaction rate.

2.3.5 RMSF analysis

RMSF analysis allows the calculation of the macromolecule flexibility during a MD simulation and B-Factor can be estimated from the RMSF values (see section 1.9.4). C_α RMSF was calculated for the 114-ns explicit water MD simulation on 55 short contiguous segments of the trajectory (2 ns each, discarding the first and the last one) and compared with the B-Factor extracted from the 1ATK PDB structure (see figure 2.7).

The calculated B-Factor of two polypeptide segments have a remarkable difference with respect to the experimental values (residues 55-70 and 85-105, colored in magenta and cyan in figure 2.3, respectively). In the 1ATK crystal, where the asymmetric unit contains only one macromolecule, the two aforementioned polypeptide segments are involved in crystal contacts and therefore their lower experimental B-Factor could be an artifact. Notably, the exosite used for docking includes several residues of the 85-105 polypeptide segment (see the mainchain colored in cyan of figure 2.3). An important *caveat* of this virtual screening procedure is that the conformations of the exosite present in solution could be very different from the one of the X-ray structure, possibly similar to the projected structures of NMA and PCA. Nevertheless, if the X-ray conformation is also present in solution, a small molecule could dock the protein in this particular exosite and therefore stabilize the X-ray conformation.

2.3.6 RMSD analysis

RMSD analysis shows the conformational stability of the protein during the MD simulation (see section 1.9.4). C_α RMSD was calculated for the 114-ns explicit water MD simulation (see figure 2.8).

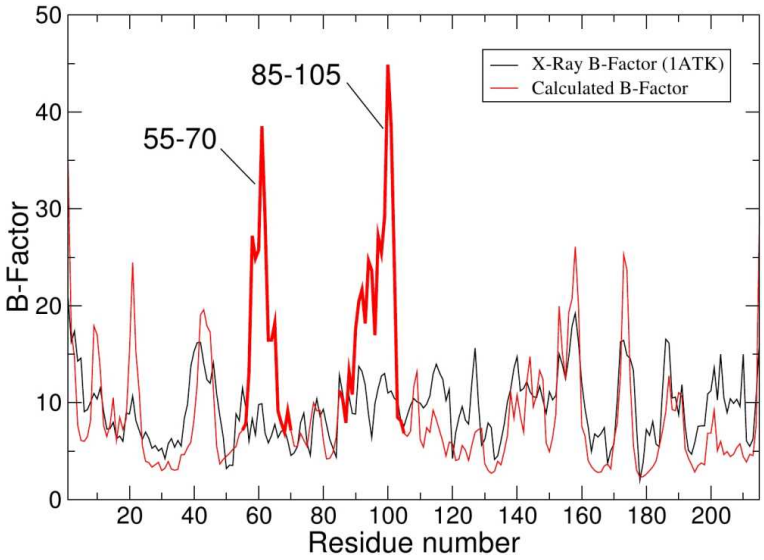


Figure 2.7: B-Factor comparison plot of cathepsin K. The C_α B-Factors extracted from the 1ATK PDB structure are in black, while the calculated B-Factor from the MD simulation are in red. The B-Factors calculated from the simulation are similar to the experimental ones. Only two polypeptide segments (residues 55-70 and 85-105, colored in magenta and cyan in figure 2.3, respectively) show a strong difference (higher B-Factor) with respect to the experimental ones. The aforementioned two segments are involved in crystal contacts and therefore their B-Factor could be artificially lower.

2.4 Virtual Screening

For docking to the Cathepsin K, two different programs suites were used: the in-house DAIM/SEED/FFLD (see section 1.7) and the commercial FRED (OpenEye Scientific Software).

DAIM/SEED/FFLD

Default parameters were used for DAIM/SEED/FFLD, except for the dielectric constant used in SEED was 2.0.

FRED

FRED is a software from OpenEye Scientific Software and is was shown to be among the best docking programs in an exhaustive study of virtual

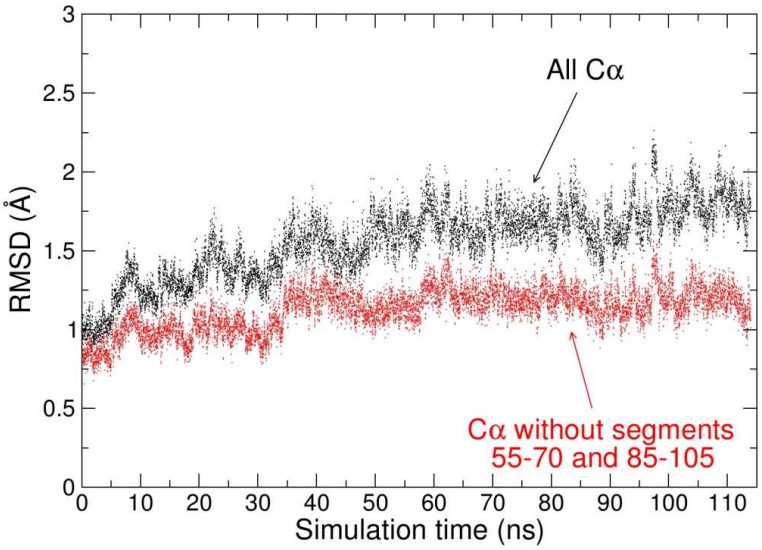


Figure 2.8: Conformational stability of cathepsin K. Cathepsin K is conformationally stable in the 114-ns MD simulation because the RMSD is in a 1.8–2.0 Å range, with respect to the energy-minimized conformation that entered in the simulation. If the two very flexible polypeptide segments identified with RMSF analysis (see figure 2.7) are not considered in the calculations, the RMSD decreases and is stable in a 1.0–1.3 Å range.

screening[61]. FRED exhaustively and rigidly docks previously generated ligand conformations into the active site of a protein. The ligand conformations are generated by OMEGA (OpenEye), which is very effective at reproducing bioactive conformations, and partial charges assigned by QuacPac (OpenEye). The docking is exhaustive, meaning that all the possible rigid rotations and traslations of a conformer are tried. The main differences of FRED with respect to the in-house procedure is that FRED considers the ligands rigid during docking and the flexibility of the ligand is introduced by pre-generating several conformations; moreover, FRED docks the entire ligand and does not require at least three fragmens as FFLD.

Default parameters were used for docking: the exhaustive search is performed with the scoring function ChemGauss3 (OpenEye) and all the poses are then ranked by consensus scoring with three different empirical scoring function: PLP[78], ChemGauss3 and OEchemscore (a slightly modified version of ChemScore[22]).

2.5 Preparation of the compound library

Given the small size of the putative binding pocket, a library of small molecules was selected for docking. It is indeed ambitious to search for low molecular weight binders, for they would very likely have a binding constant in the μM to mM range.

The ZINC database[44] subset "clean-fragments" was downloaded from the ZINC webpage[1]. This subset follows the so called *rule of three*[10] (molecular weight less than 250 Da, five or less rotatable bonds, calculated water/octanol partition coefficient less or equal to 2.5) and it is devoid of reactive and not desirable compounds, such as halides, acid anhydrides, peroxides, iso(thio)cyanates, sulphate esthers, nitrogroups, epoxides, N-O and N-N bonds and Michael acceptors[2].

CHARMm atom types[62] and partial charges were successfully assigned to 153884 of 170968 compounds (86027 neutral and 84941 charged) of the subset by the program Witnotp (A. Widmer, Novartis Pharma, unpublished). The remaining 17084 compounds, for which Witnotp failed to assign the atom types, were discarded for further calculations. A 10% failure ratio for the atom type assignment is in the average of our docking projects. Only 114425 compounds, that after fragmentation with DAIM had three or more fragments, were docked with the in-house procedure.

A total of 39459 compounds, with less than three fragments, were clustered for similarity with DAIM (leader algorithm) and the 1351 cluster representatives were docked with SEED. Since SEED rigidly docks molecules to the receptor, several different conformations were generated for every compound by an *ad hoc* procedure, mainly built on the software "CORINA"[3], to provide some degrees of flexibility. CORINA is a commercial software that generates 3D structures from a 2D representation in a automatic, reliable and robust way. The procedure follows these steps:

- Ten conformers are generated by CORINA from SMILES representations
- The CORINA generated conformers are minimized and redundant conformations are discarded (RMSD <= 0.5 Angstrom or GSEAL 3D similarity >= 0.95)

[1] http://zinc.docking.org

[2] The full list of the SMILES based rules for the definition of reactive compounds are listed in http://blaster.docking.org/filtering/rules_yuck.txt

[3] http://www.molecular-networks.com

- For every remaining conformer, ten additional conformers are generated by randomly changing torsional angles with the program "obrotamer" of the OpenBabel suite (http://openbabel.org)
- All the generated conformers are then minimized with the CHARMm and the TAFF force field[15] and redundant poses discarded. The minimization with two different force fields provides more variability in the minimized conformations.

An average of 5.18±3.16 conformers were generated for the 1351 cluster representatives mentioned above (minimum=1; maximum=15; mode=2; median=4; 1st quartil=3; 3rd quartil=7).

2.5.1 Docking to the Cathepsin K

A total of 114425 molecules with three or more fragments were docked with the classical SEED/FFLD procedure, while smaller molecules with less than three fragments were docked with SEED only.

Docking with SEED only

The normal docking procedure discards the informations about the position of all the fragments in favor of their geometrical centers, while the docking with SEED only keeps all the docked positions. For every one of the 1351 molecules docked by SEED, multiple conformers were docked and the representatives of the five energetically most favorable clusters were minimized with CHARMM and the CHARMm force field[62], using a dielectric constant of 2r.

2.5.2 Scoring

A total of 14729 poses (i.e., about 11 poses per compound) were generated by the docking with SEED only, and 1152983 poses (i.e., about 10 poses per compound) were generated by SEED/FFLD.

Cut-offs

Prior to scoring, some cut-offs on van der Waals energy and Coulombic energy were applied to weed out molecules which are unlikely to bind. The value of the cut-offs was based on the mode of the energy distribution (see figure 2.9), and poses with a van der Waals energy and a Coulombic energy more favorable than -18 kcal/mole (-2 kcal/mole for the compounds docked

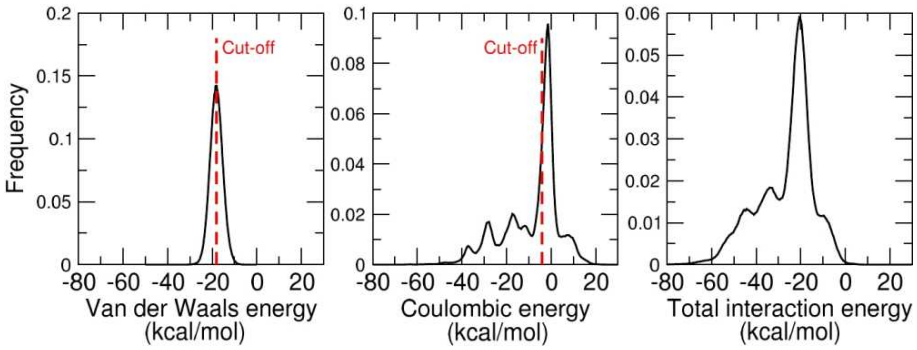


Figure 2.9: Distribution of the binding energies. The red dashed lines show the value of the cut-offs on van der Waals and Coulombic energies.

with SEED only) and -4 kcal/mole (0 kcal/mole for the compounds docked with SEED only), respectively, were kept. Additionally, only poses with two or more hydrogen bonds to the receptor were considered for scoring. The number of hydrogen bonds was calculated with CHARMM employing a cut-off distance between the heavy atoms of the donor and the acceptor of 2.6 Å and a cut-off angle (donor—H$\cdots$acceptor) of 90 degrees.

A set diagram showing the yield after applying the cut-offs is shown in figure 2.10. A total of 36904 poses of the SEED/FFLD procedure and 1111 poses of the SEED only procedure survived all the cut-offs.

SEED/FFLD

The scoring of the docking with SEED/FFLD was based on the van der Waals and coulombic interaction energies calculated by CHARMm after minimization. A novel procedure was used for separating molecules before scoring. Compounds were initially divided into two main groups, charged and neutral molecules, and subsequently into several subgroups depending on the number of rotatable bonds (from one to five) and charged molecules were also divided into smaller groups depending on their net formal charge (from plus one to plus three). The grouping were established to compare molecules with similar properties like charge and rotatable bonds. It's important to notice that the number of rotatable bonds correlates with the molecular weight (a molecule needs to be big to accommodate many rotatable bonds), but molecular weight was not considered for grouping because it is a continuous variable and therefore any grouping division would have been completely arbitrary. Moreover, the number of rotatable bonds intu-

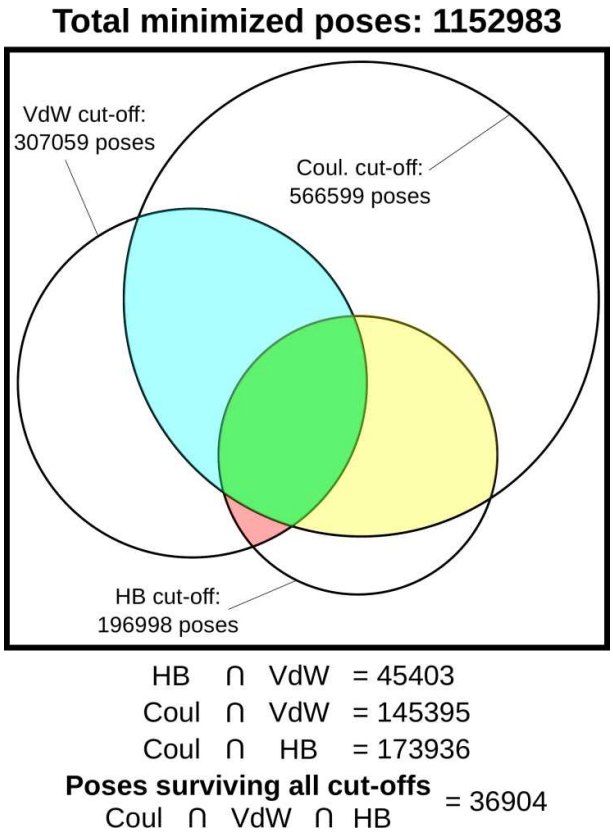


Figure 2.10: Only the 3% of the poses survives all cut-offs. The plot refers to the cut-offs applied to the results of the SEED/FFLD calculations. The surfaces of the sets (the cut-off on coulombic energy, van der Waals energy and the number H-bonds made) and of the universal set (all the minimized poses) are proportional to the number of poses, while intersections are only approximatively proportional.

itively correlates with the entropy which is lost upon binding, therefore the entropy loss contribution was implicitly taken into account by comparing molecules with the same number of rotatable bonds. For every one of the aforementioned groups, two descriptors were calculated: the total interaction energy (defined as the sum of van der Waals and coulombic interaction energies), and the ligand efficiency[39], computed as the total interaction energy divided by the molecular weight. Poses in every group were ranked according to the two descriptors and, for each applied descriptor, the most favourable twenty poses and the most favourable twenty compounds were

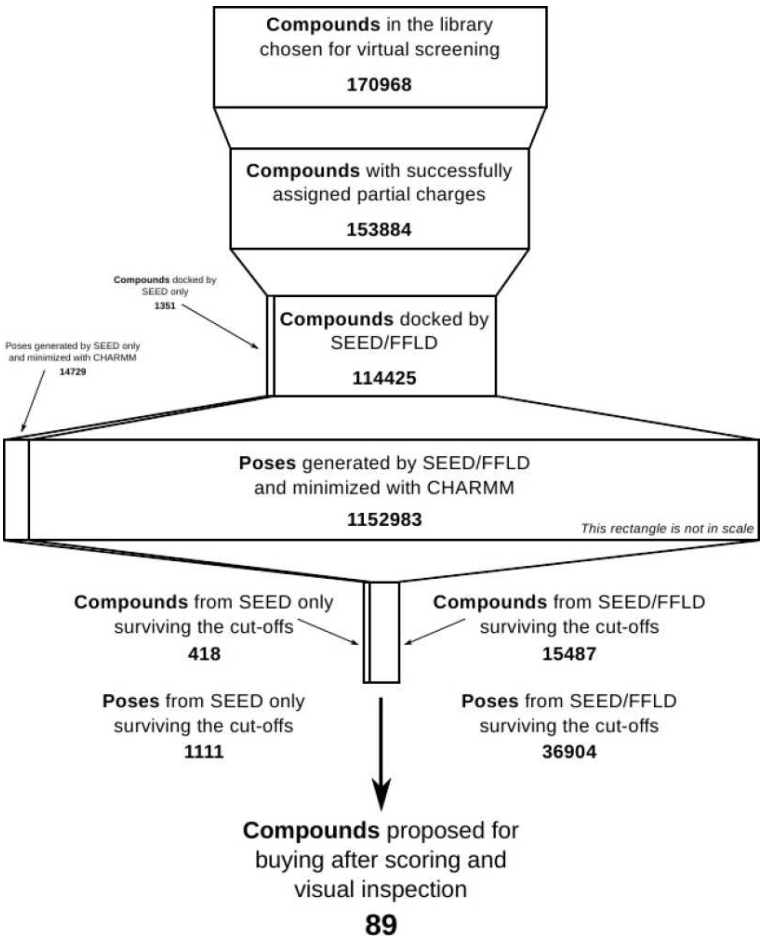


Figure 2.11: Docking acts like a filter. Docking applied to virtual screening can be considered as a filter for weeding out compounds which are unlikely to bind to the target. Applied to the Cathepsin K exosite, docking allowed to discard 170879 compounds from a library and retaining only 89 compounds, 11 of them experimentally tested.

visual inspected[4]. After visual inspection, 89 compounds were proposed for buying.

[4]In some rankings, more than one pose per compound was present in the first twenty positions.

FRED

The scoring of FRED is by default a consensus scoring based on three different scoring function: PLP[78], ChemGauss3 and OEchemscore (a slightly modified version of [22]).

2.6 Experimental procedures

The activity of Cathepsin K in the presence or absence of 1 mM compound was measured using the substrate Z-Phe-Arg-AMC (Bachem) at a final concentration of 20 μM. All experiments were performed in 50 mM HEPES buffer pH 7.4 containing 150 mM NaCl, 1 mM EDTA and 2.5 mM DTT at 25 °C. Reactions were started by adding the enzyme (final concentration 0.1 nM) into the reaction mixture containing all other components and reaction progress monitored continuously at $\lambda_{excitation}$ 383 nm and $\lambda_{emission}$ 455 nm in an Aminco SPF-500 fluorimeter.

2.7 Results

A first virtual screening was performed with SEED/FFLD on a library of 114425 compounds and with SEED only on a library of 1318 compounds. After scoring, 89 compounds were selected and 11 were experimentally tested.
One of them (ZINC16955291, figure 2.12, left), a deoxythymidine analogue, slightly activated Cathepsin K. Its binding mode showed one hydrogen bond between an hydroxy group and the carboxy group of D94; and a second hydrogen bond between the NH group the thymine and the carbonyl of the backbone of M97 (see figure 2.14, left, and figure 2.13). Interestingly, docking of the active compound with FRED showed a more reasonable alternative binding mode (see figure 2.14, middle). This binding mode was not obtained by SEED/FFLD (compare figure 2.14, left and middle, and figure 2.13, right).
A substructure search in the previously used database was performed with DAIM using a 2,6-dioxopiperidine scaffold (see figure 2.12) and yielded 716 compounds similar to the first binder. They were docked with FRED, which was preferred to SEED/FFLD because it lacks the limitation of the three fragments, and therefore it is more suitable for docking small molecules than SEED/FFLD. A total of 55 poses were selected from the docking results, 6 were experimentally tested and all of them activated Cathepsin K. The most active of them them was ZINC13282342 (see figure 2.12).

2.8 Crystal contacts, ligand binding, and cocrystallization

As mentioned in paragraph 2.3.5, the exosite is involved in crystal contacts in the crystal structure 1ATK, where the asymmetric unit contains only one monomer. Besides the aforementioned considerations on the existence of the exosite in solution, a problem might arise if the cocrystal of Cathepsin K and the second binder is obtained by soaking a previously obtained crystal of Cathepsin K in a solution of the second binder. In fact, if the binding surface of the exosite is buried by the other macromolecule in the crystal contact, the binder would never be able to diffuse and bind to the exosite. Fortunately the surface of the exosite is not completely buried and the binder should be able to bind to it (see figure 2.16, right).

In the 3HD7 structure (Cathepsin K complexed by Glycosaminoglycans), where the asymmetric unit contains two monomers, a crystal contact in the asymmetric unit is observed, involving the same exosite surface. Also in this structure, the volume between the macromolecules is enough to host the activator. Moreover, the aforementioned volume between the two macromolecules is occupied by several crystallization waters.

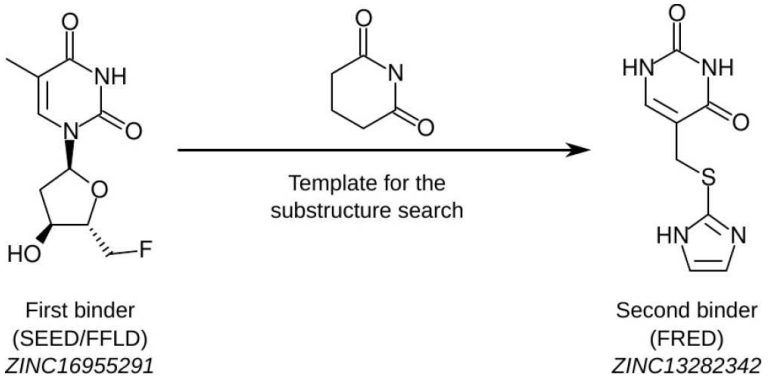


Figure 2.12: Binders found through virtual screening. A first binder (ZINC16955291, left) was found by docking with SEED/FFLD (see figure 2.14, left). Redocking this compound with FRED produced an alternative binding mode (see figure 2.14, middle). Docking with FRED a subset of similar structures resulted in a second binder (ZINC13282342, left and see figure 2.14, right).

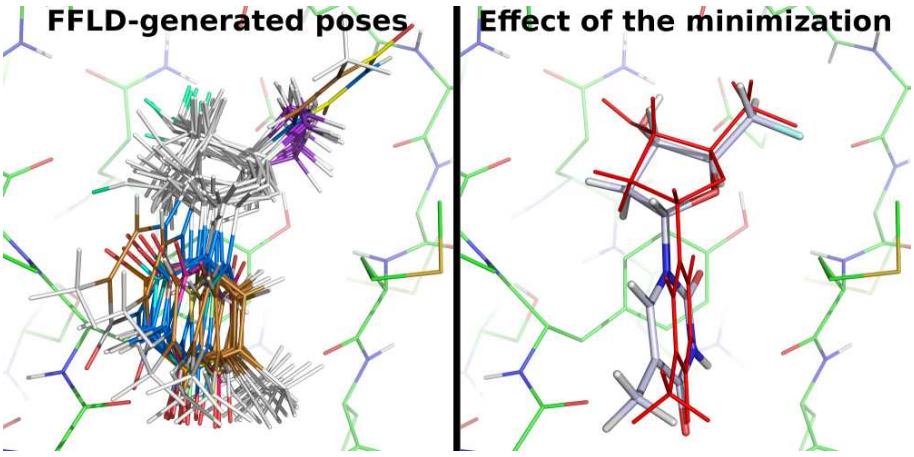


Figure 2.13: Poses of the first binder generated by FFLD and effects of the minimization. The first binder (ZINC16955291, left) was found by virtual screening with SEED/FFLD. *Left.* A total of 17 poses were generated. After minimization with CHARMM, the minimized poses were compared for similarity and 3 of them were discarded, because they fell in the same energy minimum of others and therefore their coordinates were the same of other poses. *Right.* The pose of the first binder, that has been accepted by visual inspection and proposed for buying, generated by FFLD (red) and subsequently minimized by CHARMM, is shown. The heavy-atom RMSD between the pose generated by FFLD and the minimized one is only 0.840 Å, therefore the genetic algorithm of FFLD performed well at finding a conformational and positional energy minimum.

2.9 Conclusions

Virtual screening targeting a putative allosteric pocket of Cathepsin K produced a series of compounds that bind the protein and behave as activators. A preferred thymine scaffold was found and compounds derived from it also activated Cathepsin K. The physiological role of nitrogenous bases and nucleosides in the regolation of the activity of Cathepsin K and other cysteine peptidases, such as Cathepsin B, V and L, is not known and has still to be investigated. The determination of the X-ray structure of the complex between Cathepsin K and the second binder (ZINC13282342) is also planned. A possible explanation for the activation of the enzyme by the two binders is that the movements predicted by NMA and observed in the PCA, which are thought to interfere with the active site conformation and therefore with substrate binding, are hindered.

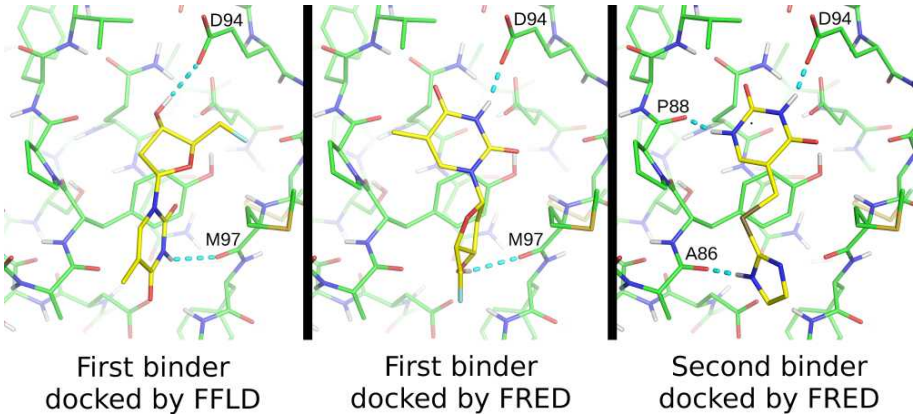


Figure 2.14: Binding modes of the two binders. The first binder (ZINC16955291, left) was found by virtual screening with SEED/FFLD. Interestingly, redocking it with FRED proposed an alternative binding mode (middle). Therefore a small subset of similar molecules was docked with FRED and a second binder was found (ZINC13282342, right).

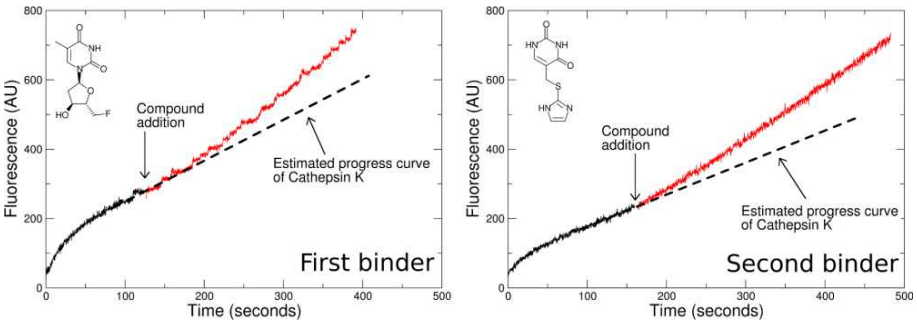


Figure 2.15: Progress curves. Progress curve of the first binder (left) and of the second binder (right). The substrate used in kinetic measurements is fluorogenic dipeptide which contains a quencher. Upon the cleavage of the peptide bond, the quencher is released and fluorescence is measured.

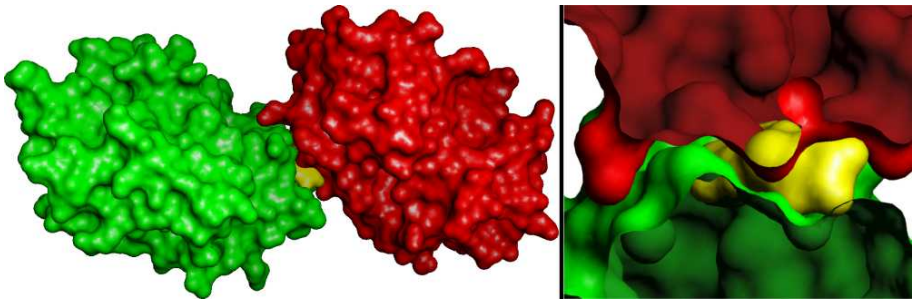


Figure 2.16: The crystal contact does not completely bury the exosite surface. The surface of the exosite is involved in crystal contacts (left). The solvent accessible surface of the original protein is in green and the one of the interaction partner obtained from the crystal lattice is in red. Fortunately, the exosite surface is not completely buried, therefore the binder (yellow surface) could diffuse and bind to it (right).

Chapter 3

A double-headed cathepsin B inhibitor devoid of warhead

Patricia Schenker, Pietro Alfarano, Peter Kolb, Amedeo Caflisch and Antonio Baici.
Protein Science, 17:2145-2155. 2008.

A double-headed cathepsin B inhibitor devoid of warhead

PATRICIA SCHENKER, PIETRO ALFARANO, PETER KOLB, AMEDEO CAFLISCH, AND ANTONIO BAICI

Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

(RECEIVED July 2, 2008; FINAL REVISION August 21, 2008; ACCEPTED August 21, 2008)

Abstract

Most synthetic inhibitors of peptidases have been targeted to the active site for inhibiting catalysis through reversible competition with the substrate or by covalent modification of catalytic groups. Cathepsin B is unique among the cysteine peptidase for the presence of a flexible segment, known as the occluding loop, which can block the primed subsites of the substrate binding cleft. With the occluding loop in the open conformation cathepsin B acts as an endopeptidase, and it acts as an exopeptidase when the loop is closed. We have targeted the occluding loop of human cathepsin B at its surface, outside the catalytic center, using a high-throughput docking procedure. The aim was to identify inhibitors that would interact with the occluding loop thereby modulating enzyme activity without the help of chemical warheads against catalytic residues. From a large library of compounds, the in silico approach identified [2-[2-(2,4-dioxo-1,3-thiazolidin-3-yl)ethylamino]-2-oxoethyl] 2-(furan-2-carbonylamino) acetate, which fulfills the working hypothesis. This molecule possesses two distinct binding moieties and behaves as a reversible, double-headed competitive inhibitor of cathepsin B by excluding synthetic and protein substrates from the active center. The kinetic mechanism of inhibition suggests that the occluding loop is stabilized in its closed conformation, mainly by hydrogen bonds with the inhibitor, thus decreasing endoproteolytic activity of the enzyme. Furthermore, the dioxothiazolidine head of the compound sterically hinders binding of the C-terminal residue of substrates resulting in inhibition of the exopeptidase activity of cathepsin B in a physiopathologically relevant pH range.

Keywords: cysteine peptidases; inhibition; enzyme kinetics; occluding loop; docking; endopeptidase; exopeptidase

Supplemental material: see www.proteinscience.org

Cathepsin B, a cysteine peptidase of the papain family (EC 3.4.22.1, identifier C01.060 in the Merops database) (Rawlings et al. 2004), has been classically ranked among the lysosomal enzymes and implicated in intracellular protein digestion. Physiologically, cathepsin B is also involved in antigen processing (Matsunaga et al. 1993), in the activation of thyroglobulin, the precursor of thyroid hormones (Friedrichs et al. 2003), and in the maturation of beta-galactosidase (Okamura-Oho et al. 1997). From a pathological point of view, cathepsin B activates trypsinogen in hereditary pancreatitis (Kukor et al. 2002) and participates in apoptosis (Bröker et al. 2005), tumor progression and malignancy (Yan and Sloane 2003; Mohamed and Sloane 2006), and rheumatic diseases (Lenarcic et al. 1988; Baici et al. 1995a,b). Particular extralysosomal functions of cathepsin B are due to altered expression at the gene level and/or atypical trafficking (Müntener et al. 2003, 2004; Zwicky et al. 2003; Baici et al. 2006).

Reprint requests to: Antonio Baici, Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland; e-mail: abaici@bioc.uzh.ch; fax: 41-44-6356805.

Abbreviations: Abz, ortho-aminobenzoyl; AMC, 7-amino-4-methylcoumarin; DTT, dithiothreitol; Z, benzyloxycarbonyl; Dnp, Nε-2,4-dinitrophenyl; FRET, Förster resonance energy transfer; DOFA, [2-[2-(2,4-dioxo-1,3-thiazolidin-3-yl)ethylamino]-2-oxoethyl] 2-(furan-2-carbonylamino)acetate; MALDI, matrix-assisted laser desorption/ionization; MS, mass spectrometry; LC, liquid chromatography; SDS-PAGE, sodium dodecyl sulfate, polyacrylamide gel electrophoresis.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.037341.108.

Cathepsin B is capable of endopeptidase (Mort 2004), peptidyl-dipeptidase (Aronson Jr. and Barrett 1978; Bond and Barrett 1980), and carboxypeptidase activities (Takahashi et al. 1986; Rowan et al. 1993). Among the cysteine peptidases, it owns a unique structural element, called an occluding loop, which comprises residues Ile^{105}–Pro^{126} (Fig. 1; Musil et al. 1991). At low pH, two salt bridges, His^{110}–Asp^{22} and Arg^{116}–Asp^{224}, hold the loop in a closed position over the primed subsites of the substrate binding cleft, thus preventing extended binding of polypeptides and endoproteolytic activity. The closed conformation, through the engagement of His^{111} in a hydrogen bond with the C-terminal carboxylate of the substrate and the loose specificity in P2′, is responsible for the peptidyl-dipeptidase activity of cathepsin B (Illy et al. 1997; Quraishi et al. 1999; Krupa et al. 2002). As mutations of the amino acids His^{110} and Asp^{22} have shown, removal of the salt bridges induces endopeptidase activity, attributed to increased flexibility of the loop (Nägler et al. 1997). This concept agrees with higher endopeptidase activity and with the competition between the occluding loop and the propeptide following deprotonation of His^{110} changing the pH from 4.0 to 6.0 (Quraishi et al. 1999). The flexibility of the loop was further demonstrated by the fact that cystatins A and C were able to displace the loop (Nycander et al. 1998; Pavlova et al. 2000), and deletion of residues 108–119 abolished the exopeptidase activity of cathepsin B (Illy et al. 1997).

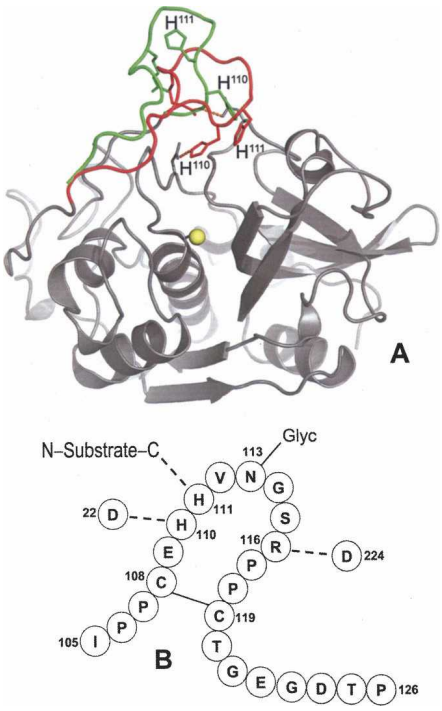


Figure 1. The occluding loop of human cathepsin B. (*A*) Closed conformation (red) in the mature enzyme (Musil et al. 1991), where a hydrogen bond is made between His^{110} and Asp^{22}. The conformation seen in procathepsin B with the occluding loop lifted (green) to accommodate the propeptide, which is not shown for clarity (Turk et al. 1996). Image generated with PyMOL software (http://www.pymol.org). (*B*) Scheme of the occluding loop with hydrogen bonds present in the closed conformation. The symbol Glyc attached to Asn^{113} indicates the N-linked oligosaccharide.

A large number of synthetic compounds behave as inhibitors of cathepsin B. Most of them are either reversible or irreversible competitive inhibitors acting ''inside the active center'' (Otto and Schirmeister 1997; Michaud and Gour 1998; Frlan and Gobec 2006). The practical use of these inhibitors is in most cases difficult for reasons analyzed elsewhere (Baici 1998). In the present study we target cathepsin B ''from the outside,'' i.e., at the surface of the molecule excluding any direct interference with catalysis. Our strategy aims at proving the concept that latching the occluding loop in its closed conformation may possibly hinder the endo- and/or exopeptidase activities of the enzyme. The approach consists of computational analysis by docking a large number of compounds to the surface of cathepsin B in the occluding loop region. From 40 compounds matching the working hypothesis, 29 are commercially available and one of them inhibits at low micromolar concentration the endo- and exoproteolytic activities of cathepsin B. We analyze the kinetic mechanism of inhibition, from which we propose a model for the interaction between the compound and the enzyme.

Results

Kinetic mechanism of inhibition with synthetic substrates

Z-RR↓AMC and Abz-GIVR↓AK(Dnp)-OH were used as substrates for monitoring the endo- and exoproteolytic activity of cathepsin B, respectively. Arrows in the substrate acronyms indicate the scissile bonds. Forty of 47,878 compounds screened in the docking approach were selected as candidate binders of human cathepsin B. Twenty-nine of them were commercially available and were tested as modifiers of enzyme activity, and one of them, DOFA (Fig. 2), behaved as a true inhibitor. Two compounds out of 29 were scarcely soluble in DMSO (Zinc-2005 codes 1271923 and 2990182; Supplemental Table 1). Five of them were soluble in DMSO but were prone to aggregation under the assay conditions for cathepsin B giving rise to modest and inconsistent inhibition between different assays. We considered these five compounds (958753, 1837733, 1871954, 3247330, and 1581881) as promiscuous inhibitors (Shoichet 2006). None of the compounds tested was a covalent inactivator of cathepsin B. DOFA did not inhibit human cathepsin L, and papain was only partially inhibited at millimolar concentrations (assays with Z-FR↓AMC; data not shown). A

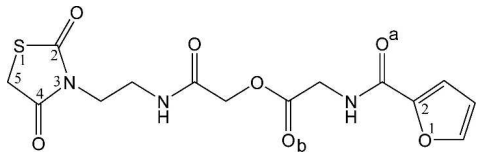


Figure 2. Structure of the inhibitor. The IUPAC name of the compound is [2-[2-(2,4-dioxo-1,3-thiazolidin-3-yl)ethylamino]-2-oxoethyl] 2-(furan-2-carbonylamino)acetate (abbreviated DOFA; Zinc-2005 code: 2616818). Numbering and lettering used as aid for the description in the main text.

previous report of high-throughput screening classified DOFA inactive against cathepsin B (http://pubchem.ncbi.nlm.nih.gov/, CID 2078629, AID 453). Reasons can be sought in either too strict criteria or in the occasional failure of high-throughput screening methods in detecting hits (Buxser and Vroegop 2005). DOFA behaved as a reversible inhibitor of cathepsin B and manifested neither tight-binding nor slow-binding inhibition behavior. Reaction traces were linear from the very beginning of the reaction, i.e., from 3–4 ms onward, as measured with a stopped-flow apparatus. Therefore, initial velocities at variable substrate and modifier concentrations were treated as steady-state rates. In control experiments we checked the possibility that any inhibitory effect of DOFA was not due to aggregation of the compound. For this purpose 0.01% Triton X-100 was added to buffers and the DTT concentration was increased to 5 mM. In comparison with buffer not containing detergent and with DTT = 2 mM, the basal activity of cathepsin B was higher by ~40%. However, inhibition profiles for increasing DOFA concentration, percentages of inhibition, and inhibition constants were the same.

The kinetic results in this study refer to those obtained with the enzyme purified from human liver. Nevertheless, we confirmed inhibition mechanisms and kinetic parameters also with recombinant cathepsin B. The reason was to ascertain any influence of glycosylation at N^{113}, which is located at the tip of the occluding loop, on the binding of potential inhibitors. Both kinetic and modeling results discussed below suggest that N^{113} glycosylation possibly occurring in the wild-type enzyme does not affect inhibition of cathepsin B by DOFA.

The kinetic mechanism was analyzed by a combination of graphical and regression methods to discriminate between models of enzyme–modifier interaction (Fig. 3). With the Abz-GIVR↓AK(Dnp)-OH exoproteolytic substrate at pH 4.5, the best fitting model for cathepsin B inhibition by DOFA was linear competitive inhibition with $I_{90/10} = 81$ and $K_i = 6.7$ μM (Fig. 4A). The specific velocity plot, typical for this mechanism, is shown in the inset of Figure 4A. With this substrate, data at pH 6.0 could not be obtained because of its scarce solubility. DOFA inhibited the endoproteolytic cathepsin B activity (Z-RR↓AMC as substrate) at pH 4.5 and 6.0. The specific velocity plot (not shown) established the linear nature of the inhibition

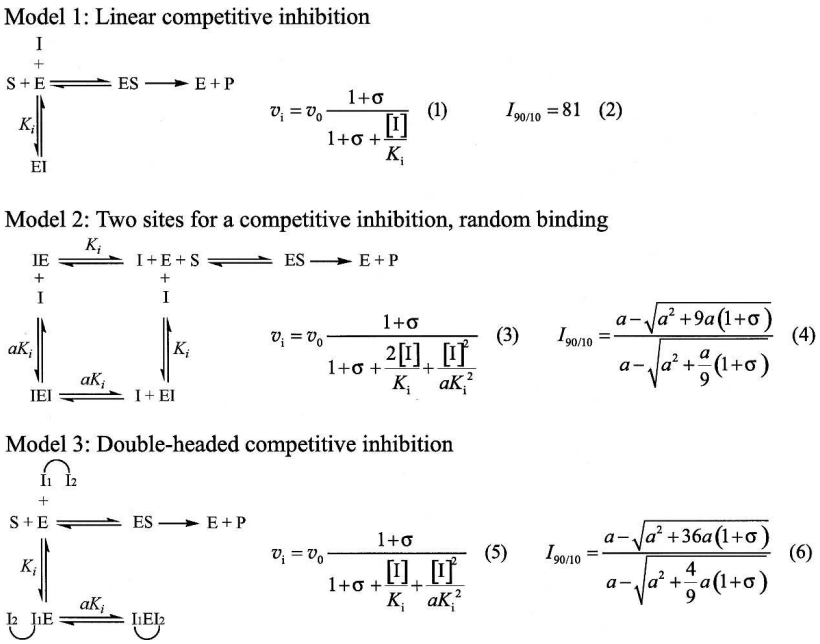


Figure 3. Kinetic models for the inhibition of cathepsin B. v_i and v_0 represent reaction rates in the presence and in the absence of inhibitor, respectively. $\sigma = [S]/K_m$; $I_{90/10}$ = ratio of the inhibitor concentrations, which give 90% and 10% inhibition. The double-headed inhibitor possesses two distinct binding moieties that sequentially bind the enzyme.

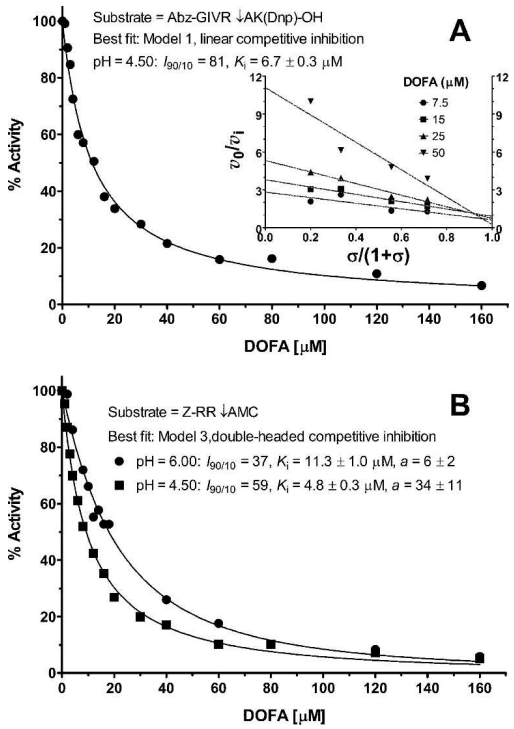


Figure 4. Inhibition profiles of human cathepsin B by DOFA. Main conditions and best-fit kinetic parameters are shown. (*A*) Inhibition profile using the FRET substrate Abz-GIVR↓AK(Dnp)-OH (pH 4.5), [S] = K_m = 9.7 μM; data were obtained fluorimetrically with $\lambda_{ex}/\lambda_{em}$ at 320/420 nm. The *inset* shows the specific velocity plot at four substrate and four inhibitor concentrations. (*B*) Inhibition profiles with Z-RR↓AMC as substrate; [S] = K_m = 0.47 mM at pH 6.0 (black circles) and [S] = K_m = 1.54 mM at pH 4.5 (black squares). Data at pH 6.0 were collected photometrically at 360 nm and those at pH 4.5 were obtained fluorimetrically with $\lambda_{ex}/\lambda_{em}$ at 383/455 nm. Fluorescence readings in *A* and *B* were corrected for the inner filter effect.

process, i.e., enzyme activity was driven to zero at saturating inhibitor concentration, and diagnosed competitive-type inhibition. However, the specific velocity plot would not be able to discriminate between Models 1–3 in Figure 3. The activity profiles obtained at a fixed substrate concentration and variable inhibitor concentrations revealed the $I_{90/10}$ ratio to be less than 81, which can be estimated by inspection to be around 60 at pH 4.5 and around 40 at pH 6.0 (Fig. 4B). This suggests that the inhibition mechanism was not of the classical competitive type, for which $I_{90/10} = 81$ (Model 1 in Fig. 3). On the other hand, Models 2 and 3 in the same figure predict $I_{90/10}$ values less than 81. For discriminating between models with Z-RR↓AMC as substrate, we set its concentration equal to K_m so that $\sigma = [S]/K_m = 1$ in the activity profiles in Figure 4B (thereby it was imperative to carefully measure the substrate concentration and K_m). In Models 2 and 3, when the factor $a = 1$, $I_{90/10}$ depends only on σ, whereas for $a \neq 1$ the $I_{90/10}$ ratio depends both on a and σ (Fig. 5A,B). Model 2 was ruled out as redundant, and Model 3 was preferred on the basis of the following quantitative and logical observations. In Model 2, two inhibitor molecules should bind to different places at the surface of the cathepsin B molecule with almost the same affinity, i.e., with the a value between 1 and 2, as shown in Figure 5A. The double-headed competitive inhibitor (Model 3 in Fig. 3) was superior also on the basis of the following quantitative observations. For Model 3, the squared symbol with dashed arrows in Figure 5B indicates the best-fit value for $a = 6$ at pH 6.0, which corresponds to $I_{90/10} = 37$ and $K_i = 11.3$ μM (Fig. 4B). Similarly, at pH 4.5, Model 3, with $a = 34$ (the black circle in Fig. 5B), $I_{90/10} = 59$, and $K_i = 4.8$ μM, produced a better fit to experimental data (Fig. 4B) than Model 2.

Inhibition of endo- and exoproteolysis using protein substrates

While low molecular mass oligopeptides are indispensable tools for determining inhibition mechanisms, they do not represent typical substrates for evaluating true endo- and exoproteolytic activities. Therefore, inhibition of cathepsin B endo- and exoproteolysis by DOFA was further

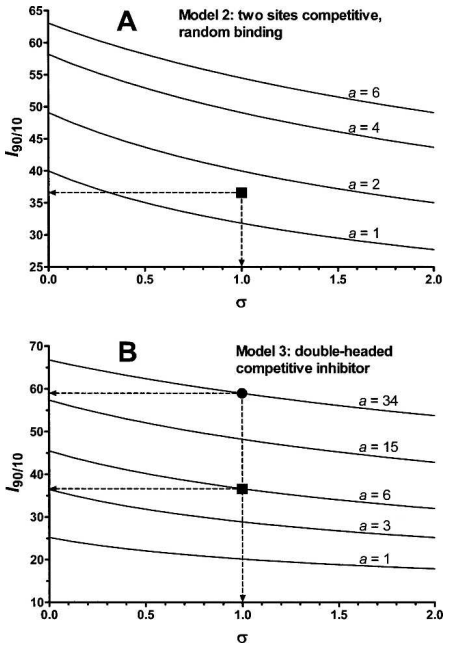


Figure 5. Ratios of inhibitor concentrations that give 90% and 10% inhibition. For Models 2 and 3 in Figure 3 the $I_{90/10}$ ratio depends on a and σ = [S]/K_m. Curves were generated with Equations 4 and 6 shown in Figure 3. (*A*) Curves for Model 2; (*B*) curves for Model 3. The black squares in panels *A* and *B* correspond to the best-fit value of the coefficient a, which coincides with the value obtained by inspection of the inhibition profiles at pH 6.0 in Figure 4B. The black circle in panel *B* shows the best-fit value of the coefficient a (34), close to the value obtained by inspection of the inhibition profiles at pH 4.5 in Figure 4B.

investigated with two protein substrates: rabbit muscle aldolase and the oxidized β-chain of bovine insulin. Peptide-bond hydrolysis on both proteins at pH 6.0 was inhibited in a concentration-dependent manner by DOFA (Fig. 6) and was almost complete at saturating inhibitor concentration. As commonly observed when switching from oligopeptide to polypeptide substrates, the modifier concentrations needed for achieving inhibition were higher.

Cathepsin B has been previously reported to sequentially remove up to nine dipeptide units from the C terminus of aldolase with peptidyl-dipeptidase activity (Bond and Barrett 1980). Aldolase incubated with cathepsin B in the presence or absence of DOFA was subjected to SDS-PAGE under reducing conditions. The bands excised after electrophoretic separation were N-terminally sequenced and the masses of the polypeptides determined by mass spectrometry. The SDS gels shown in Figure 7 were purposely overloaded to reveal the less prominent bands numbered 2, 3, and 4. The N-terminal sequence of the polypeptides in bands 1 and 2 was PHSHPAL, which corresponds to positions 2–8 of aldolase. Mass analysis of the whole broad band 1 revealed the presence, besides intact aldolase with amino acids 2–364 and a mass of 39,216 Da, C-terminally degraded fragments corresponding to sequences 2–362, 2–360, and other masses down to 2–352 that were generated by the sequential removal of six dipeptide units from the C terminus. The significant presence of the 2–351 peptide indicates that aldolase degradation also comprises carboxypeptidase activity of cathepsin B. The tiny band 2 contained the polypeptide 2–347, which was the smallest fragment obtained under the described conditions and corresponded to the removal of 17 amino acids from the C terminus, i.e., 8 dipeptide units (peptidyl–dipeptidase activity) plus one amino acid (carboxypeptidase activity). Bands 3 and 4 had the common N terminus SIGTENT (positions 46–52), suggesting their generation by a minor (but very useful for the purposes of this test) endoproteolytic cleavage between Gln[45] and Ser[46]. Band 3

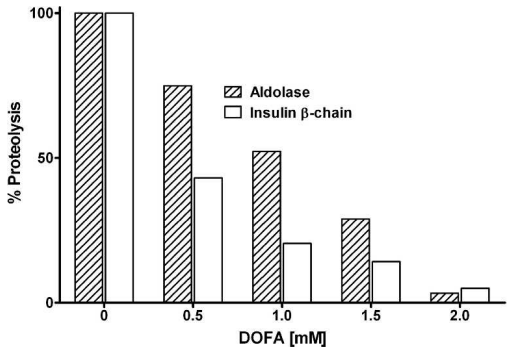


Figure 6. Inhibition of cleavage of rabbit muscle fructose 1,6-bisphosphate aldolase and of the bovine insulin β-chain by cathepsin B. Concentration-dependent inhibition by DOFA at pH 6.0. Released peptides were quantified fluorimetrically by the fluorescamine method.

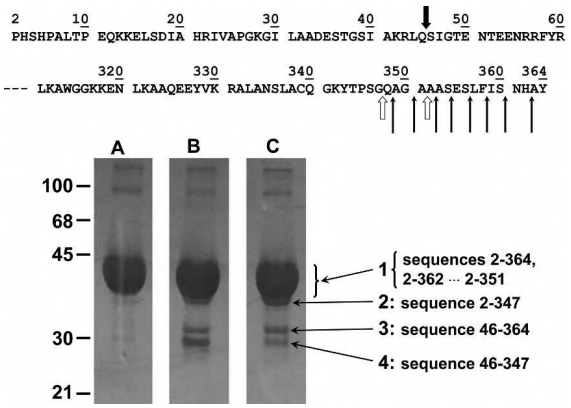


Figure 7. Proteolysis of rabbit muscle fructose 1,6-bisphosphate aldolase by cathepsin B. Aldolase was incubated 4 h at 25°C with cathepsin B at pH 6.0, and products were separated by SDS-PAGE. (Lane *A*) Aldolase alone run in parallel as control; (lane *B*) aldolase plus cathepsin B; (lane *C*) aldolase with cathepsin B and DOFA. The N- and C-terminal parts of the rabbit aldolase sequence, with numbered amino acid residues, are shown at the *top* of the figure. Thin arrows indicate sequential peptidyl-dipeptidase cleavages by cathepsin B. The two open thick arrows show carboxypeptidase cleavages. The black thick arrow indicates a minor endoproteolytic attack site. Numbers on the *left* side indicate the molecular masses of markers in kilodaltons. Numbers 1–4 with long pointing arrows are assigned to the bands for referencing in the main text, and the peptides identified within the bands are shown next to the band numbers. Peptides were identified in the excised bands by N-terminal sequencing and MALDI analysis.

corresponded to the 46–364 sequence and band 4 corresponded to sequence 46–347. Comparison of the relative intensities of bands 3 and 4 in lanes B and C (Fig. 7) suggests that DOFA inhibited the exopeptidase activity of cathepsin B on the macromolecular substrate aldolase, while the inhibitor was less or not active against the endoproteolytic activity of cathepsin B, as shown by the persistence of band 3 in lane C.

The bovine insulin β-chain is a helpful substrate for exploring the specificity of peptide-bond cleavage by peptidases, including human cathepsin B (McKay et al. 1983). The cleavage at multiple sites of the insulin β-chain by cathepsin B was almost completely inhibited by a saturating concentration of DOFA at pH 6.0 (Fig. 6). Samples of insulin β-chain incubated with cathepsin B in the presence and absence of DOFA were separated by liquid chromatography, and the eluted peaks were analyzed by mass spectrometry. We confirmed most of the cathepsin B cleavages previously described (McKay et al. 1983). Additional cleavages, which depended on the incubation time and the relative enzyme to substrate concentrations, were identified. Each of the chromatographic peaks in Figure 8 contained only one (e.g., peptides 1–13, 1–14, and 6–28) or more proteolytic fragments (e.g., peptides 20–30 plus 1–11). Only a few of them will be discussed to demonstrate the inhibitory activity of DOFA. The generation of peptides 1–11 and 20–30 by endoproteolysis was inhibited by

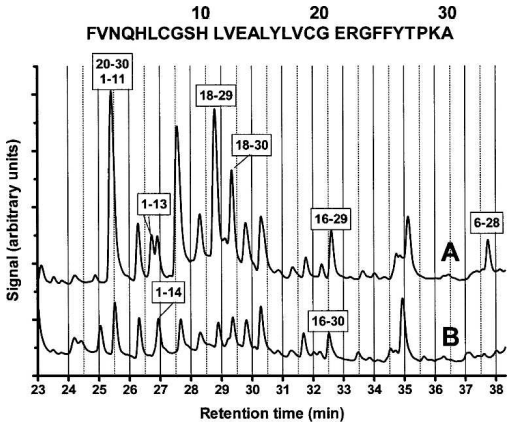


Figure 8. Proteolysis of the bovine insulin β-chain by cathepsin B. Incubation at pH 5.5 and 25°C in the absence (*A*) and in the presence (*B*) of DOFA. Peptides released by enzymatic digestion were characterized by LC/MALDI/MS. The sequence at the *top* of the figure shows the sequence of bovine insulin, β-chain with amino acid numbering. Boxed numbers, e.g., 6–28, indicate peptides released by the endo- and exoproteolytic action of cathepsin B.

DOFA. Fragment 1–13 was a product of carboxypeptidase activity on peptide 1–14: While DOFA completely inhibited this exoproteolytic event, endoproteolytic cleavage between residues 14 and 15 was unaffected, as shown by comparing the chromatographic profiles in the absence (Fig. 8, profile A) and in the presence of DOFA (Fig. 8, profile B). Fragment 18–29, generated by endoproteolysis at position 17–18 and exoproteolysis at position 29–30, was strongly decreased in the presence of DOFA. Comparison of the areas of fragments 18–29 and 18–30 after inhibition shows that both the exo- and the endoproteolytic events concerned with them were affected by the inhibitor. Carboxypeptidase activity inhibition is also shown by comparing the disappearance of fragment 16–29 and the appearance of fragment 16–30. Finally, the peptidyl–dipeptidase cleavage shown by fragment 6–28 disappeared in the presence of DOFA. Since the 6–30 peptide in Figure 8, profile B could not be unambiguously assigned, inhibition could be ascribed to endoproteolytic cleavage at positions 5–6, peptidyl–dipeptidase inhibition at position 28–29, or both.

Discussion

The docking approach used in this study aimed at identifying compounds that bind to the occluding loop of human cathepsin B and thereby modulate its enzymatic activity. Considering that no other means are available for selecting small molecules for a rather flexible protein segment, the most diverse compound library was used. There is a fundamental difference between our methodology and previous approaches to inhibitors that might interact with parts of the occluding loop of cathepsin B (Murata et al. 1991; Michaud and Gour 1998; Cathers et al. 2002). These studies exploited the properties of the fungal inhibitor E-64 (Hanada et al. 1978) and the crystal structure of cathepsin B (Musil et al. 1991). New inhibitors, such as CA-074, were designed by modification of the E-64 structure to create interactions with the primed sites of the substrate binding cleft and thus with the occluding loop (Murata et al. 1991). In this approach the epoxy moiety of the modifiers acted as a warhead against the active site cysteine to produce potent irreversible inhibition (Matsumoto et al. 1999). Conversely, we did not target the active site, and the molecules selected by docking were devoid of chemically active warheads. The finding of only one active compound out of 29 substances is in line with previous studies, which showed that candidate inhibitors predicted by docking algorithms are not necessarily active in vitro (Huang et al. 2005).

In agreement with published data (Illy et al. 1997; Nägler et al. 1997; Quraishi et al. 1999; Krupa et al. 2002) our results support the notion that the occluding loop of cathepsin B is a highly flexible segment. Short episodes of opening and closing govern endo- and exoproteolysis, which thus occur intermittently in the pH range 4.5–6.0. The compound identified by docking, DOFA, behaves as a reversible competitive inhibitor of the exo- and endopeptidase activities of cathepsin B with no involvement of a substrate-like binding mode in the active center of the enzyme. It can inhibit endoproteolysis most likely by forcing the loop in the closed conformation, and exoproteolysis by disturbing the accommodation of the C terminus of polypeptides in the primed side of the binding cleft. Docking calculations suggest that the dioxothiazolidine moiety of DOFA is located in the binding pocket of the side chain of the substrate C-terminal residue when the loop is closed. It makes a hydrogen bond with $N\delta^1$ of His^{110} and sterically hinders substrate binding for exoproteolytic activity (Fig. 9A,B). In this figure, the structures of DOFA and the four amino acids VRAK, which are part of the FRET substrate used for measuring exoproteolytic activity, were modeled on the known structure of human cathepsin B merely to show their spatial relationship. In fact, such a ternary ESI complex does not exist according to the kinetic mechanism; only the ES and EI complexes exist. This structural arrangement is consistent with the kinetic mechanism of inhibition: At pH 4.5 the exoproteolytic activity of cathepsin B is inhibited in a linear competitive manner by virtue of the dioxothiazolidine ring. Docking suggests that DOFA binds at the surface of the occluding loop and is involved in three hydrogen bonds with the enzyme (Fig. 9A). Of particular interest is that the carbonyl oxygen involved in a hydrogen bond with the amide of His^{111} occupies the place of the oxygen of a conserved water molecule present in the crystal structures of human cathepsin B. This position is shown as a red sphere in Figure 9B. Together with the favorable van der Waals interactions, three hydrogen bonds stabilize the loop in

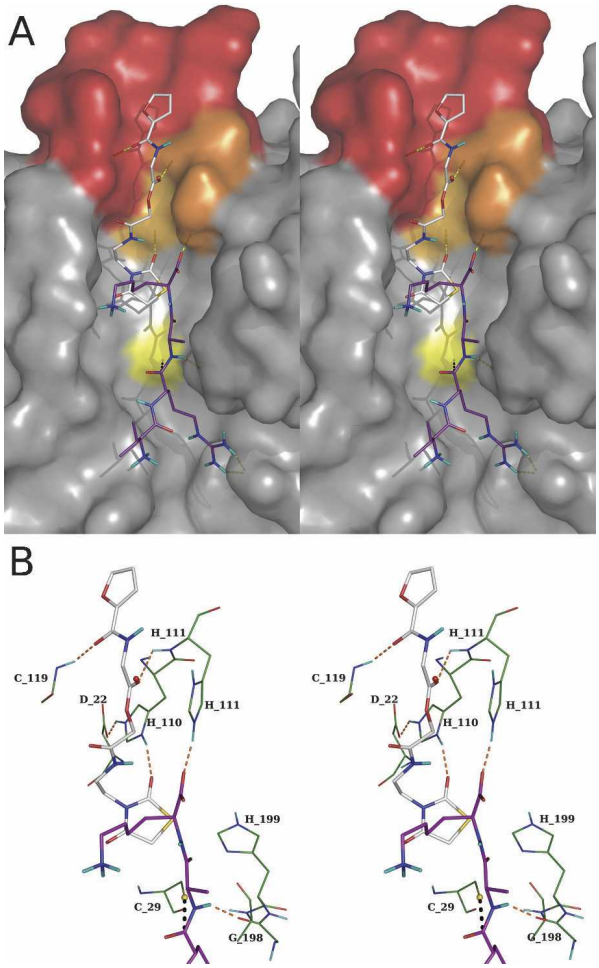


Figure 9. Stereoview of inhibitor and substrate within the structure of human cathepsin B. (*A*) The protein surface is shown with the occluding loop in red, His^{110} and His^{111} in orange, and the catalytic Cys^{29} in yellow. The substrate (carbon atoms in purple) is shown as residues VR↓AK, which is part of the FRET substrate used in this study to monitor peptidyl–dipeptidase activity. The Dnp group, which is exposed to the solvent, was omitted for clarity. The pose of DOFA obtained by docking (carbon atoms in gray) is visible in the *upper* part with the furan moiety in the occluding loop and the dioxothiazolidine ring sterically interfering with the lysine side chain of the substrate. (*B*) Details of the presumed interactions between DOFA and cathepsin B residues (carbon atoms in green). The inhibitor makes three hydrogen bonds with residues of the occluding loop: carbonyl group labeled "a" in Figure 2 with amide nitrogen of Cys^{119}, carbonyl group labeled "b" in Figure 2 with amide nitrogen of His^{111}, and 2-oxo group of the dioxothiazolidine ring with $N\delta^1$ of His^{110}. The hydrogen bonds of the occluding loop in the closed conformation, His^{110}–Asp^{22} and His^{111} with the C-terminal carbonyl of the substrate, are also shown. The red sphere indicates the position of a water molecule in the cathepsin B crystal structure. Images generated with PyMOL software (see Fig. 1).

the closed conformation (Fig. 9B). These structural properties are reflected in the kinetic mechanism observed with Z-RR↓AMC. Despite not being a complete model for endoproteolysis, the dipeptide is unquestionably cleaved by cathepsin B in the endo-, not in the exoproteolytic mode. The double-headed competitive kinetic mechanism diagnosed for DOFA on Z-RR↓AMC, and the changes in the kinetic parameters observed when going from pH 4.5 to 6.0, insinuate cooperation between the furan and the dioxothiazolidine moieties. Kinetic measurements alone would be unable to identify which element makes the stronger contribution to binding energy. However, with the support of structural information from modeling discussed above, we suggest that the dioxothiazolidine moiety takes the leader function. In the double-headed mechanism at pH 4.5, the K_i value of 4.8 μM is thus ascribed to the dioxothiazolidine part of the molecule and $K_i \approx 160$ μM to the furan part (factor $a = 34$ in Model 3, Fig. 3). The increase of pH to 6.0 disfavors the K_i of the dioxothiazolidine portion, which becomes 11.3 μM and favors binding of the furan moiety, decreasing its K_i to 68 μM.

The measurements with protein substrates, aldolase and the insulin β-chain, confirm the results just discussed with synthetic substrates, as well as the previously recorded ability of cathepsin B to exert peptidyl–dipeptidase and carboxypeptidase activity in the pH range 5.5–6.0 (Mort et al. 1998). While efficiently inhibiting these two exoproteolytic activities, DOFA was not able to inhibit all endoproteolytic events. For instance, the minor endoproteolytic action at position 45–46 of rabbit muscle aldolase was only moderately or not affected by DOFA, as was not position 14–15 of the insulin β-chain. Yet, other endoproteolytic attacks on the insulin β-chain were efficiently inhibited.

The pooled experimental evidence from the above described properties of DOFA inhibition strengthens the notion that not only pH dependent, intermittent closing and opening of the occluding loop, but also substrate structure determine together the susceptibility of particular peptide bonds. The binding affinity of DOFA is possibly not high enough to compete with a substrate, which optimally binds at both unprimed and primed sites. It appears that the occluding loop of cathepsin B can be assisted in assuming its closed or open conformation by particular polypeptides with an induced fit mechanism that promotes either exo- or endoproteolysis. This observation forms the basis of cathepsin B inhibition by its own propeptide (Quraishi et al. 1999) and its resistance to cystatin A and C inhibition (Nycander et al. 1998; Pavlova et al. 2000).

The above results draw attention to the physiopathological role of cathepsin B, which exerts its action in a relatively broad range of pH values, with exoproteolysis in the range 5–6 and endoproteolysis in the range 5–7.4. The pH in early endosomes is ~6.2 and drops to 5.0–5.5 in late endosomes and lysosomes (Gu and Gruenberg 2000). Extracellularly, the pH can vary from relatively acidic, e.g., at the periphery of solid tumors (Rofstad et al. 2006), to 7.4 in blood. Contrary to commonplace textbook information, proteolysis within phagolysosomes does not quite occur in a very acidic environment. For instance, 5 min after initiation of

phagocytosis in human polymorphonuclear leukocytes, the relatively acidic milieu of the lysosomes (pH ≈ 5) increases in the phagocytic vacuoles to a value of 7.8. It falls then to 7.4 after 15 min, to 6.4 after 30 min, and 5.7 after 60 min (rounded values from Cech and Lehrer [1984]). Various hydrolases are expected to find their optimal conditions of action during this forth and back excursion of pH (Baici et al. 1996).

It has been suggested that "targeting of the occluding loop of cathepsin B may be a poor inhibitor design strategy if the enzyme environment has a pH greater than 5.5" (Cathers et al. 2002). This statement refers to the above discussed approach "from inside." Since endo- and exoproteolysis by cathepsin B occurs in the physiologically and pathologically relevant pH range of 5–6, our approach of targeting the occluding loop from the enzyme surface may represent an alternative starting point for the development of clinically active compounds.

Materials and Methods

Materials and general methods

Recombinant human cathepsin B was prepared from *Escherichia coli* inclusion bodies (Kuhelj et al. 1995). Wild-type cathepsin B from human liver was obtained from Calbiochem. The synthetic substrates Z-FR↓AMC[2] and Z-RR↓AMC were purchased from Bachem and the internally quenched fluorogenic substrate Abz-GIVR↓AK(Dnp)OH from Merck. Rabbit muscle fructose-1,6-bisphosphate aldolase (EC 4.1.2.13) and fluorescamine were from Sigma-Aldrich Chemie and the oxidized bovine insulin β-chain from Serva Feinbiochemica. The putative binders/modifiers of cathepsin B activity determined by docking analysis and their sources are listed in Supplemental Table 1. The compound studied in detail, [2-[2-(2,4-dioxo-1,3-thiazolidin-3-yl)ethylamino]-2-oxoethyl] 2-(furan-2-carbonylamino)acetate (IUPAC name, here abbreviated DOFA; Fig. 2), Zinc-2005 code 2616818 (Irwin and Shoichet 2005) was obtained from Enamine. The compounds were dissolved as concentrated stock solutions in dimethyl sulphoxide, stored at 4°C, and diluted into the appropriate buffer at the moment of the assay. Photometric measurements were carried out with a Cary50 UV-visible spectrophotometer, and fluorescence was measured with an Aminco SPF-500 fluorimeter operating in the ratio mode. Reversed phase HPLC was performed in a Hewlett Packard Series 1100 apparatus with a Nucleosil 120-5 C18 column (4.0 × 250 mm). Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) and liquid chromatography followed by mass analysis, LC-MALDI/MS, was performed at the Functional Genomics Center in Zurich.

Docking

The A and B chains of the human cathepsin B structure, entry 1HUC in the Brookhaven database, were considered and all water molecules were removed. Hydrogens were added to side chains and termini of the protein to simulate conditions at pH 6–7. Cys^{29} was in the reduced form, and His^{110} was protonated at both imidazole nitrogen atoms. CHARMm atom types and force field parameters were assigned and used for both protein and ligands (Momany and Rone 1992), and hydrogens were minimized with the program CHARMM (Brooks et al. 1983). The search for a putative binding site focused on the occluding loop of the enzyme, for which no inhibitor has been reported to date. A benzene molecule was used as a probe with the program SEED (Solvation Energy for Exhaustive Docking) to identify possible binding pockets whose main interaction component is supposed to be hydrophobicity (Majeux et al. 1999). The surface tested by SEED consists of cathepsin B residues within a 9 Å distance from His^{110}, which locks the occluding loop in a closed position. The benzenes with the most favorable free energy of binding, excluding those in the catalytic site, were used to define the binding site for docking, which included all the 42 residues of the protein that had >50% of their atoms within 5 Å from the optimal poses of benzenes.

About 3 million compounds in the ZINC 2005 library (Irwin and Shoichet 2005) were clustered using a 17-field fingerprint by Decomposition and Identification of Molecules (Kolb and Caflisch 2006). Molecules with more than nine rotatable bonds were neglected. The resulting 48,026 cluster representatives were docked as in previous in silico campaigns (Huang et al. 2005, 2006; Kolb et al. 2008). For each molecule docked by the program FFLD (Fragment-based Flexible Ligand Docking) (Budin et al. 2001; Cecchini et al. 2004), the 300 most favorable poses were clustered with the leader algorithm yielding 764,776 poses (~17 poses per compound, 44,527 compounds). The cluster representatives were minimized in the rigid protein using CHARMM.

It is necessary to weed out unlikely poses to prevent a high number of false positives (Kolb et al. 2008). Suitable filters are cutoffs in the total van der Waals interaction energy and in the van der Waals efficiency (defined as the ratio between van der Waals energy and molecular mass). A cutoff of −30 kcal/mol for intermolecular van der Waals energy was selected by plotting a histogram of the van der Waals energies (Supplemental Fig. 1). Analogously, a cutoff of −0.09 kcal/g was chosen for the van der Waals efficiency (Supplemental Fig. 2). A total of 106,502 poses of 10,709 compounds survived both filters. Poses closer than 4 Å to Cys^{29} of the catalytic site and those with no hydrogen bond to the protein were neglected, yielding 25,178 remaining poses (5678 compounds).

The electrostatic contribution to the binding free energy was evaluated by the finite difference Poisson approach (module PBEQ of CHARMM) (Im et al. 1998). Poses were ranked for visual inspection according to most favorable van der Waals energy, van der Waals efficiency, electrostatics, and sum of van der Waals and electrostatic energy. Poses with high ranking in two lists or more were selected and visually inspected. Forty poses belonging to 40 different compounds were selected in this way. According to the predicted binding modes, three regions of the protein were exploited for binding: one on the occluding loop and one each on immediately adjacent regions. Twenty-nine out of the 40 candidate compounds were commercially available.

Kinetics

The peptide Z-RR↓AMC with 7-amino-4-methlycoumarin (AMC) as leaving group is a handy substrate for measuring the endopeptidase activity of cathepsin B (Barrett 1980), in which the two arginine residues correspond to the P_1 and P_2 positions (Schechter and Berger 1967). We preferred this substrate over Z-FR↓AMC for its slower hydrolysis that allowed long measuring times with low substrate consumption. The buffer for measurements at pH 6.00 was 50 mM sodium phosphate

containing 2 mM EDTA and 2 mM dithiothreitol (DTT), hereafter referred to as pH 6.0 buffer. For measurements at pH 4.50, 0.1 M sodium acetate, 0.2 M NaCl, 2 mM EDTA, and 2.5 mM DTT were used (hereafter referred to as pH 4.5 buffer). In control experiments aimed at ascertaining any aggregation phenomena on the part of the putative inhibitors, the DTT concentration was increased to 5 mM and the buffer included 0.01% (v/v) of hydrogenated Triton X-100, which does not absorb light in the ultraviolet range and is not fluorescent. The buffers were prepared and used at the same temperature of the kinetic assays, 25 ± 1°C. The final concentration of dimethyl sulphoxide, used for preparing stock solutions of the compounds, was 0.2% (v/v) in experiments with synthetic substrates and 1% (v/v) in those with protein substrates. Reaction progress with Z-RR↓AMC at pH 4.5 and 6.0 was followed by either measuring the change in absorbance at 360 nm, with $\Delta\varepsilon = 11{,}400\ M^{-1}cm^{-1}$ as the difference absorption coefficient between free and bound AMC or fluorimetrically with excitation and emission wavelengths, λ_{ex} and λ_{em}, set at 383 and 455 nm, respectively. Exopeptidase activity was continuously monitored fluorimetrically with the Förster resonance energy transfer (FRET) labeled substrate Abz-GIVR↓AK(Dnp)-OH (Cotrin et al. 2004). Appropriate corrections for inner filter effects were taken into account in fluorescence measurements considering the absorbencies of the substrates and of the tested compounds at λ_{ex} and λ_{em} (Palmier and Van Doren 2007). With the AMC-based substrate, λ_{ex} at 383 nm allowed high substrate concentrations to be used with small inner filter corrections. Since with the FRET substrate inner filter effects were more pronounced, λ_{ex} and λ_{em} were selected according to the compound being tested. The combination $\lambda_{ex}/\lambda_{em}$ 320/420 nm, often used in assays with this type of substrate, could not be utilized for the strong absorbance at 320 nm of one of the compounds (ZINC-2005 code 1797383), which simulated strong enzyme inhibition, whereas the combination $\lambda_{ex}/\lambda_{em}$ 370/430 nm revealed no inhibition. Any modifier-independent contribution to enzyme activity loss by denaturation, adherence to the vessel walls, or other causes was monitored in time-course assays by the Selwyn method (Selwyn 1965). Screening and primary analysis of the kinetic inhibition mechanism of the compounds was performed on initial velocity data obtained at various concentrations of substrate and compounds. The specific velocity plot, a sensitive tool for detecting mixed-type and/or hyperbolic mechanisms was routinely used in this phase (Baici 1981). The kinetic inhibition models considered and their related rate equations are shown in Figure 3. In Models 2 and 3, a is a coefficient, which is equal to 1 when the first binding process does not influence the second one. If the equilibrium constant of the vacant site is influenced after binding to the first site, $a \neq 1$. Kinetic analysis was performed by fitting the equations in Figure 3 to experimental data using GraphPad Prism version 5.00 for Windows. The runs test and analysis of residuals were performed to monitor deviations from a model, and discrimination between mechanisms was made by analysis of variance of the difference between the sum of squares (extra sum-of-squares test) and calculation of F ratios and p values. According to Occam's razor principle, the simplest mechanism describing experimental data was considered superior to a more complex (redundant) mechanism. Additionally, a powerful and decisive tool for model discrimination was the $I_{90/10}$ ratio, i.e., the ratio of inhibitor concentration necessary to achieve 90% and 10% inhibition. This parameter was determined experimentally and compared with values calculated from Equations 4 and 6 in Figure 3.

Protein substrates

The activity of cathepsin B on rabbit muscle aldolase was quantified by measuring primary amino groups derived from peptide-bond hydrolysis with fluorescamine (Schwabe 1975). Aldolase was dissolved in pH 6.0 buffer without DTT, incubated 3 h at 37°C with cathepsin B, which was pre-activated with DTT, and DOFA under vigorous shaking to prevent precipitation. Final concentrations were: aldolase = 3.5 mg/mL = 21.7 μM, cathepsin B = 2.3 μM, DOFA = 0.5, 1.0, 1.5, and 2.0 mM. The reaction was stopped with trichloroacetic acid, 5% (w/v) final concentration, and centrifuged. We added 0.1 mL of the clear supernatant to 2.0 mL 0.2 M sodium borate buffer (pH 8.50), and 1.0 mL of fluorescamine solution (15 mg/100 mL acetone). The fluorescence of labeled peptides was measured with $\lambda_{ex}/\lambda_{em}$ = 390/480 nm. Blanks were subtracted and controls compared with the samples containing the compounds. Measurements with the oxidized β-chain of bovine insulin as substrate were performed in pH 6.0 buffer using the same procedure. Final concentrations were: insulin β-chain = 72 μg/mL = 20.6 μM, cathepsin B = 2.3 μM, DOFA = 0.5, 1.0, 1.5, and 2.0 mM.

Endo- and exoproteolysis of protein substrates

Reaction products resulting from incubation of aldolase with cathepsin B and DOFA were also assessed by denaturing gel electrophoresis. Final concentrations in incubation mixtures at 25°C under continuous shaking for 4 h were: aldolase 3.5 mg/mL (21.7 μM) in pH 6.0 buffer without DTT, 2.3 μM of preactivated cathepsin B, and 1.0 mM DOFA. After centrifugation, the clear supernatant was analyzed by sodium dodecyl sulphated, polyacrylamide gel electrophoresis (SDS-PAGE) under reducing conditions with 12.5% polyacrylamide and silver stained. For N-terminal sequencing by Edman degradation, the bands were transferred to a polyvinylidene fluoride membrane. Peptides in the excised bands were also identified by MALDI analysis.

The oxidized β-chain of bovine insulin was incubated at a final concentration of 1.0 mg/mL (286 μM) in 0.1 M sodium acetate buffer containing 0.2 mM NaCl, 2 mM EDTA, and 2.5 mM DTT (pH 5.50) with 1.2 μM cathepsin B (control) or cathepsin B plus 1.0 mM DOFA. After 20 h at 25°C under shaking, solutions were centrifuged and the solvent removed in a Speed Vac concentrator. The residue was resuspended in 60 μL of deionized water containing 0.1% trifluoroacetic acid and centrifuged before analysis by LC/MALDI/MS.

Electronic supplemental material

Supplemental Table 1 shows the identification codes and the sources of the cathepsin B inhibitors. Supplemental Figure 1 presents the van der Waals interaction energies distribution used in the docking procedure and shows the cutoff of −30 kcal/mol for intermolecular van der Waals energy. Analogously, Supplemental Figure 2 shows the van der Waals interaction energy efficiencies distribution.

Acknowledgments

Docking calculations were performed on Matterhorn, a Beowulf Linux cluster at the University of Zurich, and we thank C. Bolliger, T. Steenbock, and A. Godknecht for computer support. We thank A. Widmer (Novartis Pharma, Basel, Switzerland) for providing the molecular modeling program Wit!P, which was used for preparing the structures. We also thank the staff at the

Functional Genomics Center in Zurich for expert assistance in LC/MALDI/MS measurements. This work was supported by grant 31-113345/1 of the Swiss National Foundation and by the Albert-Böni Foundation (to A.B.), and by a grant of the Hartmann-Müller Foundation (to A.C.).

References

Aronson Jr., N.N. and Barrett, A.J. 1978. The specificity of cathepsin B. Hydrolysis of glucagon at the C-terminus by a peptidyldipeptidase mechanism. *Biochem. J.* **171:** 759–765.

Baici, A. 1981. The specific velocity plot. A graphical method for determining inhibition parameters for both linear and hyperbolic enzyme inhibitors. *Eur. J. Biochem.* **119:** 9–14.

Baici, A. 1998. Inhibition of extracellular matrix-degrading endopeptidases: Problems, comments, and hypotheses. *Biol. Chem.* **379:** 1007–1018.

Baici, A., Hörler, D., Lang, A., Merlin, C., and Kissling, R. 1995a. Cathepsin B in osteoarthritis: Zonal variation of enzyme activity in human femoral head cartilage. *Ann. Rheum. Dis.* **54:** 281–288.

Baici, A., Lang, A., Hörler, D., Kissling, R., and Merlin, C. 1995b. Cathepsin B in osteoarthritis: Cytochemical and histochemical analysis of human femoral head cartilage. *Ann. Rheum. Dis.* **54:** 289–297.

Baici, A., Szedlacsek, S.E., Früh, H., and Michel, B.A. 1996. pH-Dependent hysteretic behaviour of human myeloblastin (leucocyte proteinase 3). *Biochem. J.* **317:** 901–905.

Baici, A., Müntener, K., Willimann, A., and Zwicky, R. 2006. Regulation of human cathepsin B by alternative mRNA splicing: Homeostasis, fatal errors and cell death. *Biol. Chem.* **387:** 1017–1021.

Barrett, A.J. 1980. Fluorimetric assays for cathepsin B and cathepsin H with methylcoumarylamide substrates. *Biochem. J.* **187:** 909–912.

Bond, J.S. and Barrett, A.J. 1980. Degradation of fructose-1,6-bisphosphate aldolase by cathepsin B. A further example of peptidyldipeptidase activity of this proteinase. *Biochem. J.* **189:** 17–25.

Bröker, L.E., Kruyt, F.A.E., and Giaccone, G. 2005. Cell death independent of caspases: A review. *Clin. Cancer Res.* **11:** 3155–3162.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Budin, N., Majeux, N., and Caflisch, A. 2001. Fragment-based flexible ligand docking by evolutionary optimization. *Biol. Chem.* **382:** 1365–1372.

Buxser, S. and Vroegop, S. 2005. Calculating the probability of detection for inhibitors in enzymatic or binding reactions in high-throughput screening. *Anal. Biochem.* **340:** 1–13.

Cathers, B.E., Barrett, C., Palmer, J.T., and Rydzewski, R.M. 2002. pH Dependence of inhibitors targeting the occluding loop of cathepsin B. *Bioorg. Chem.* **30:** 264–275.

Cecchini, M., Kolb, P., Majeux, N., and Caflisch, A. 2004. Automated docking of highly flexible ligands by genetic algorithms: A critical assessment. *J. Comput. Chem.* **25:** 412–422.

Cech, P. and Lehrer, R.I. 1984. Phagolysosomal pH of human neutrophils. *Blood* **63:** 88–95.

Cotrin, S.S., Puzer, L., Judice, W.A.D., Juliano, L., Carmona, A.K., and Juliano, M.A. 2004. Positional-scanning combinatorial libraries of fluorescence resonance energy transfer peptides to define substrate specificity of carboxydipeptidases: Assays with human cathepsin B. *Anal. Biochem.* **335:** 244–252.

Friedrichs, B., Tepel, C., Reinheckel, T., Deussing, J., von Figura, K., Herzog, V., Peters, C., Saftig, P., and Brix, K. 2003. Thyroid functions of mouse cathepsins B, K, and L. *J. Clin. Invest.* **111:** 1733–1745.

Frlan, R. and Gobec, S. 2006. Inhibitors of cathepsin B. *Curr. Med. Chem.* **13:** 2309–2327.

Gu, F. and Gruenberg, J. 2000. ARF1 regulates pH-dependent COP functions in the early endocytic pathway. *J. Biol. Chem.* **275:** 8154–8160.

Hanada, K., Tamai, M., Yamagishi, M., Ohmura, S., Sawada, J., and Tanaka, I. 1978. Isolation and characterization of E-64, a new thiol protease inhibitor. *Agric. Biol. Chem.* **42:** 523–528.

Huang, D.Z., Lüthi, U., Kolb, P., Edler, K., Cecchini, M., Audetat, S., Barberis, A., and Caflisch, A. 2005. Discovery of cell-permeable non-peptide inhibitors of β-secretase by high-throughput docking and continuum electrostatics calculations. *J. Med. Chem.* **48:** 5108–5111.

Huang, D.Z., Lüthi, U., Kolb, P., Cecchini, M., Barberis, A., and Caflisch, A. 2006. In silico discovery of β-secretase inhibitors. *J. Am. Chem. Soc.* **128:** 5436–5443.

Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., and Mort, J.S. 1997. Role of the occluding loop in cathepsin B activity. *J. Biol. Chem.* **272:** 1197–1202.

Im, W., Beglov, D., and Roux, B. 1998. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson–Boltzmann equation. *Comput. Phys. Commun.* **111:** 59–75.

Irwin, J.J. and Shoichet, B.K. 2005. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45:** 177–182.

Kolb, P. and Caflisch, A. 2006. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem.* **49:** 7384–7392.

Kolb, P., Huang, D., Dey, F., and Caflisch, A. 2008. Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. *J. Med. Chem.* **51:** 1179–1188.

Krupa, J.C., Hasnain, S., Nägler, D.K., Menard, R., and Mort, J.S. 2002. S2′ substrate specificity and the role of His110 and His111 in the exopeptidase activity of human cathepsin B. *Biochem. J.* **361:** 613–619.

Kuhelj, R., Dolinar, M., Pungercar, J., and Turk, V. 1995. The preparation of catalytically active human cathepsin B from its precursor expressed in *Escherichia coli* in the form of inclusion bodies. *Eur. J. Biochem.* **229:** 533–539.

Kukor, Z., Mayerle, J., Krüger, B., Tóth, M., Steed, P.M., Halangk, W., Lerch, M.M., and Sahin-Tóth, M. 2002. Presence of cathepsin B in the human pancreatic secretory pathway and its role in trypsinogen activation during hereditary pancreatitis. *J. Biol. Chem.* **277:** 21389–21396.

Lenarcic, B., Gabrijelcic, D., Rozman, B., Drobnic-Kosorok, M., and Turk, V. 1988. Human cathepsin B and cysteine proteinase inhibitors (CPIs) in inflammatory and metabolic joint diseases. *Biol. Chem. Hoppe Seyler* **369:** 257–261.

Majeux, N., Scarsi, M., Apostolakis, J., Ehrhardt, C., and Caflisch, A. 1999. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins* **37:** 88–105.

Matsumoto, K., Mizoue, K., Kitamura, K., Tse, W.C., Huber, C.P., and Ishida, T. 1999. Structural basis of inhibition of cysteine proteases by E-64 and its derivatives. *Biopolymers* **51:** 99–107.

Matsunaga, Y., Saibara, T., Kido, H., and Katunuma, N. 1993. Participation of cathepsin B in processing of antigen presentation to MHC class II. *FEBS Lett.* **324:** 325–330.

McKay, M.J., Offermann, M.K., Barrett, A.J., and Bond, J.S. 1983. Action of human liver cathepsin B on the oxidized insulin B chain. *Biochem. J.* **213:** 467–471.

Michaud, S. and Gour, B.J. 1998. Cathepsin B inhibitors as potential anti-metastatic agents. *Expert Opin. Ther. Pat.* **8:** 645–672.

Mohamed, M.M. and Sloane, B.F. 2006. Cysteine cathepsins: Multifunctional enzymes in cancer. *Nat. Rev. Cancer* **6:** 764–775.

Momany, F.A. and Rone, R. 1992. Validation of the general purpose QUANTA®3.2/CHARMm® force field. *J. Comput. Chem.* **13:** 888–900.

Mort, J.S. 2004. Cathepsin B. In *Handbook of proteolytic enzymes*, 2nd ed. (eds. A.J. Barrett and N.D. Rawlings, J.F. Woessner, Jr.), pp. 1079–1086. Elsevier, London, UK.

Mort, J.S., Magny, M.C., and Lee, E.R. 1998. Cathepsin B: An alternative protease for the generation of an aggrecan ''metalloproteinase'' cleavage neoepitope. *Biochem. J.* **335:** 491–494.

Müntener, K., Zwicky, R., Csucs, G., and Baici, A. 2003. The alternative use of exons 2 and 3 in cathepsin B mRNA controls enzyme trafficking and triggers nuclear fragmentation in human cells. *Histochem. Cell Biol.* **119:** 93–101.

Müntener, K., Zwicky, R., Csucs, G., Rohrer, J., and Baici, A. 2004. Exon skipping of cathepsin B: Mitochondrial targeting of a lysosomal peptidase provokes cell death. *J. Biol. Chem.* **279:** 41012–41017.

Murata, M., Miyashita, S., Yokoo, C., Tamai, M., Hanada, K., Hatayama, K., Towatari, T., Nikawa, T., and Katunuma, N. 1991. Novel epoxysuccinyl peptides. Selective inhibitors of cathepsin B, in vitro. *FEBS Lett.* **280:** 307–310.

Musil, D., Zucic, D., Turk, D., Engh, R.A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katunuma, N., et al. 1991. The refined 2.15 Å X-ray crystal structure of human liver cathepsin B: The structural basis for its specificity. *EMBO J.* **10:** 2321–2330.

Nägler, D.K., Storer, A.C., Portaro, F.C.V., Carmona, E., Juliano, L., and Menard, R. 1997. Major increase in endopeptidase activity of human cathepsin B upon removal of occluding loop contacts. *Biochemistry* **36:** 12608–12615.

Nycander, M., Estrada, S., Mort, J.S., Abrahamson, M., and Björk, I. 1998. Two-step mechanism of inhibition of cathepsin B by cystatin C due to displacement of the proteinase occluding loop. *FEBS Lett.* **422:** 61–64.

Okamura-Oho, Y., Zhang, S.Q., Callahan, J.W., Murata, M., Oshima, A., and Suzuki, Y. 1997. Maturation and degradation of β-galactosidase in the post-Golgi compartment are regulated by cathepsin B and a non-cysteine protease. *FEBS Lett.* **419:** 231–234.

Otto, H.H. and Schirmeister, T. 1997. Cysteine proteases and their inhibitors. *Chem. Rev.* **97:** 133–171.

Palmier, M.O. and Van Doren, S.R. 2007. Rapid determination of enzyme kinetics from fluorescence: Overcoming the inner filter effect. *Anal. Biochem.* **371:** 43–51.

Pavlova, A., Krupa, J.C., Mort, J.S., Abrahamson, M., and Björk, I. 2000. Cystatin inhibition of cathepsin B requires dislocation of the proteinase occluding loop. Demonstration by release of loop anchoring through mutation of His110. *FEBS Lett.* **487:** 156–160.

Quraishi, O., Nägler, D.K., Fox, T., Sivaraman, J., Cygler, M., Mort, J.S., and Storer, A.C. 1999. The occluding loop in cathepsin B defines the pH dependence of inhibition by its propeptide. *Biochemistry* **38:** 5017–5023.

Rawlings, N.D., Tolle, D.P., and Barrett, A.J. 2004. MEROPS: The peptidase database (http://merops.sanger.ac.uk). *Nucleic Acids Res.* **32:** D160–D164.

Rofstad, E.K., Mathiesen, B., Kindem, K., and Galappathi, K. 2006. Acidic extracellular pH promotes experimental metastasis of human melanoma cells in athymic nude mice. *Cancer Res.* **66:** 6699–6707.

Rowan, A.D., Feng, R., Konishi, Y., and Mort, J.S. 1993. Demonstration by electrospray mass spectrometry that the peptidylpeptidase activity of cathepsin B is capable of rat cathepsin B C-terminal processing. *Biochem. J.* **294:** 923–927.

Schechter, I. and Berger, A. 1967. On the size of the active sites in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27:** 157–162.

Schwabe, C. 1975. A fluorescence assay for proteolytic enzymes. *Anal. Biochem.* **53:** 484–490.

Selwyn, M.J. 1965. A simple test for inactivation of an enzyme during assay. *Biochim. Biophys. Acta* **105:** 193–195.

Shoichet, B.K. 2006. Screening in a spirit haunted world. *Drug Discov. Today* **11:** 607–615.

Takahashi, T., Dehdarani, A.H., Yonezawa, S., and Tang, J. 1986. Porcine spleen cathepsin B is an exopeptidase. *J. Biol. Chem.* **261:** 9375–9381.

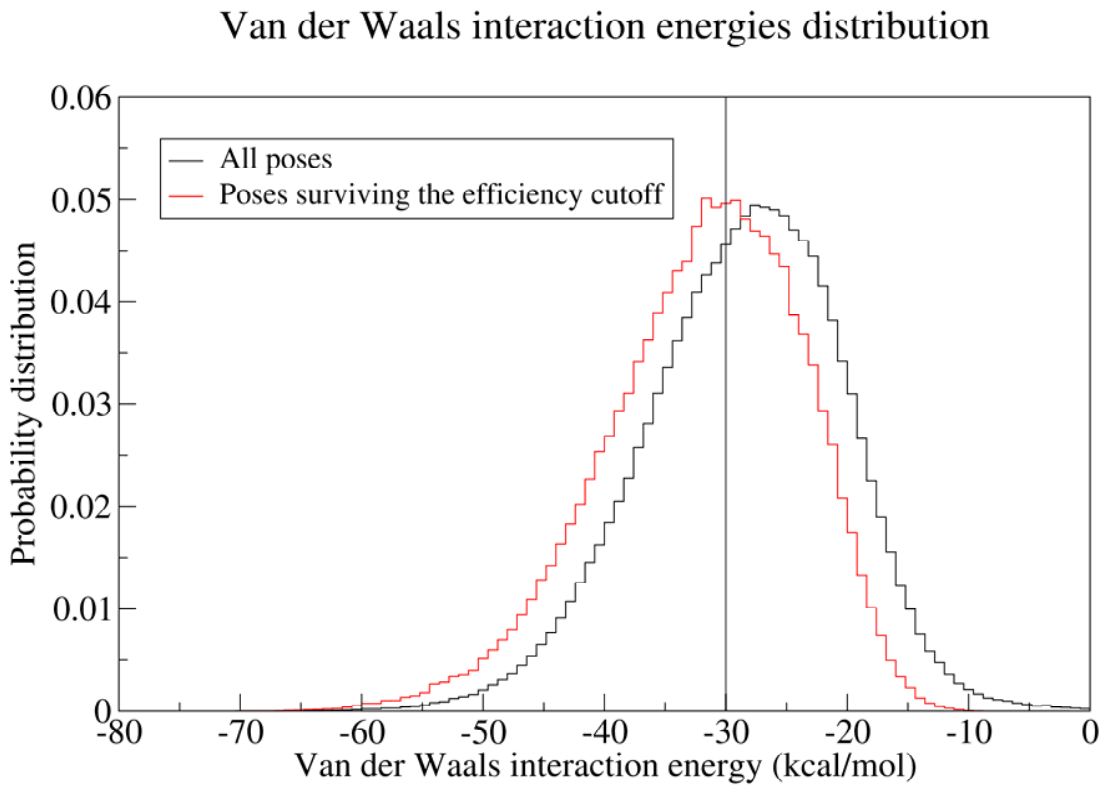
Turk, D., Podobnik, M., Kuhelj, R., Dolinar, M., and Turk, V. 1996. Crystal structures of human procathepsin B at 3.2 and 3.3 Å resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide. *FEBS Lett.* **384:** 211–214.

Yan, S.Q. and Sloane, B.F. 2003. Molecular regulation of human cathepsin B: Implication in pathologies. *Biol. Chem.* **384:** 845–854.

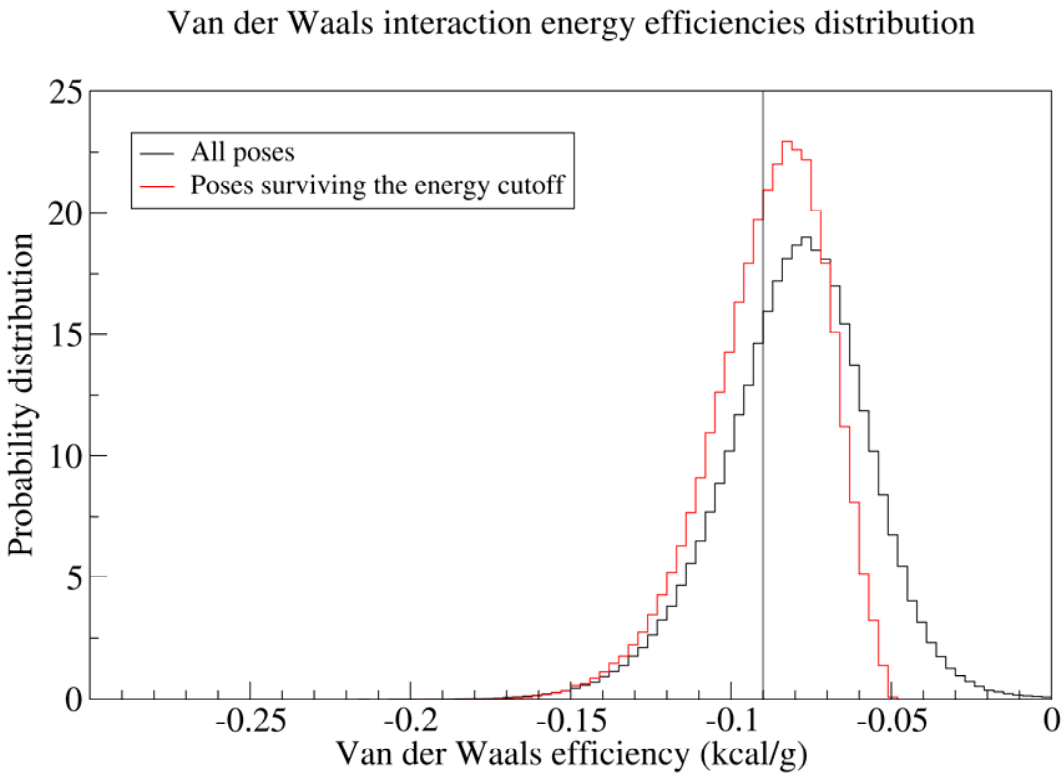
Zwicky, R., Müntener, K., Csucs, G., Goldring, M.B., and Baici, A. 2003. Exploring the role of 5′-alternative splicing and of the 3′-untranslated region of cathepsin B mRNA. *Biol. Chem.* **384:** 1007–1018.

Supplemental Table S1. Identification codes and source of 29 putative inhibitors of human cathepsin B. The ZINC-codes are those of the database 2005 [Irwin, J. J., and Shoichet, B. K. (2005) J. Chem. Inf. Model. 45, 177-182]. [2-(2,4-dioxothiazolidin-3-yl) ethylcarbamoylmethyl 2-(furan-2-carbonylamino) acetate (DOFA) is shown in boldface.

Source	ZINC 2005-codes
ChemDiv (San Diego, USA)	33816, 978038, 2487954, 2560859
Ambinter (Paris, France)	1271923, 1414169
ChemBridge (San Diego, USA)	958753, 2904166, 2973678, 2990182
Pharmerks (Moscow, Russia)	989395, 1755678, 1837733, 1871954
Life Chemicals (Burlington, Canada)	1797383
Enamine (Kiev, Ukraine)	**2616818**, 2667966, 3247330, 3333255, 3340535, 3417832, 3439551, 3451184
IBScreen (Moscow, Russia)	1419203, 2128374, 2131752
National Cancer Institute, U.S.A.	1581881, 1628238, 1682930



Supplemental Figure S1. **Van der Waals interaction energies distribution.** The histogram represents the distribution of van der Waals interaction energies before and after the efficiency cutoff. The vertical line marks the cutoff on van der Waals energy.



Supplemental Figure S2. **Van der Waals interaction energy efficiencies distribution.** The histogram represents the distribution of van der Waals interaction energy efficiencies before and after the energy cutoff. The vertical line represents the cutoff of van der Waals efficiency.

Chapter 4

The Chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 Is Essential for H3K27me3 Binding and Function during Arabidopsis Development

Vivien Exner, Ernst Aichinger, Huan Shu, Thomas Wildhaber, Pietro Alfarano, Amedeo Caflisch, Wilhelm Gruissem, Claudia Kohler, Lars Hennig.

PlosOne, 4(4): 2009.

OPEN ACCESS Freely available online

PLoS one

The Chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 Is Essential for H3K27me3 Binding and Function during Arabidopsis Development

Vivien Exner[1], Ernst Aichinger[1], Huan Shu[1], Thomas Wildhaber[1], Pietro Alfarano[2], Amedeo Caflisch[2], Wilhelm Gruissem[1], Claudia Köhler[1], Lars Hennig[1]*

1 Department of Biology & Zurich-Basel Plant Science Center, ETH Zurich, Zurich, Switzerland, **2** Department of Biochemistry, University of Zurich, Zurich, Switzerland

Abstract

Polycomb group (PcG) proteins are essential to maintain gene expression patterns during development. Transcriptional repression by PcG proteins involves trimethylation of H3K27 (H3K27me3) by Polycomb Repressive Complex 2 (PRC2) in animals and plants. PRC1 binds to H3K27me3 and is required for transcriptional repression in animals, but in plants PRC1-like activities have remained elusive. One candidate protein that could be involved in PRC1-like functions in plants is LIKE HETEROCHROMATIN PROTEIN 1 (LHP1), because LHP1 associates with genes marked by H3K27me3 *in vivo* and has a chromodomain that binds H3K27me3 *in vitro*. Here, we show that disruption of the chromodomain of *Arabidopsis thaliana* LHP1 abolishes H3K27me3 recognition, releases gene silencing and causes similar phenotypic alterations as transcriptional *lhp1* null mutants. Therefore, binding to H3K27me3 is essential for LHP1 protein function.

Citation: Exner V, Aichinger E, Shu H, Wildhaber T, Alfarano P, et al. (2009) The Chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 Is Essential for H3K27me3 Binding and Function during Arabidopsis Development. PLoS ONE 4(4): e5335. doi:10.1371/journal.pone.0005335

Editor: Hany A. El-Shemy, Cairo University, Egypt

Received February 3, 2009; **Accepted** March 22, 2009; **Published** April 28, 2009

Funding: This work was supported by SNF (www.snf.ch) grant 3100AO-116060 (to L.H.), by FP6 IP AGRON-OMICS (ec.europa.eu) contract LSHG-CT-2006-037704 (to W.G.) and by ETH projects (www.ethz.ch) TH-16/05-2 (to L.H.) and TH-12/06-1 (to C.K.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lhennig@ethz.ch

Introduction

Polycomb group (PcG) proteins maintain gene expression patterns during development in animals and plants by establishing a cellular memory system for transcriptional repression [1]. Although many functional details of PcG proteins remain unknown, current models suggest that repression involves trimethylation of histone H3 lysine 27 (H3K27me3) by Polycomb repressive complex 2 (PRC2). In insects and mammals, H3K27me3 assists in the recruitment of PRC1 [2]. Binding of PRC1 to H3K27me3 is mediated by the chromodomain of the PRC1 subunit Polycomb (Pc) [3]. Metazoan PRC1 complexes catalyze H2A monoubiquitylation via their RING-domain subunits and are needed for stable repression of PcG target genes [2]. Although the PcG system is present in plants and PRC2 homologs have similar functions, no clear plant PRC1 homologs have been identified [1]. Proteins that may have PRC1-like functions in plants include EMBRYONIC FLOWER 1, VERNALIZATION 1, LIKE HETEROCHROMATIN PROTEIN 1 (LHP1) and RAWUL-proteins [4–7].

The gene for *Arabidopsis thaliana* LHP1 was first found in screens for mutants with altered leaf glucosinolate levels and named *TU8* [8,9] as well as in screens for inflorescence meristem function and named *TERMINAL FLOWER 2* [10,11]. In addition, LHP1 was identified as a homolog of metazoan HETEROCHROMATIN PROTEIN1 (HP1) [12]. Similar to HP1, LHP1 contains a chromodomain and a chromo shadow domain [11,12]. Unlike HP1, however, LHP1 is usually localized in euchromatin and is needed for maintenance of gene silencing in euchromatin but not in heterochromatin [13,14]. Finally, LHP1 can bind to H3K27me3 *in vitro* and associates with genes marked by H3K27me3 *in vivo* [15,16]. Homologs of the animal PRC1 core component RING1 have recently been identified in Arabidopsis, and binding of AtRING1A to LHP1 suggests similar structure and function of plant and animals PRC1 complexes [17].

Together, the model has emerged that LHP1 binds to PcG target loci that have been trimethylated at H3K27 by PRC2 to establish persistent transcriptional repression. We tested this hypothesis using a LHP1 mutant with a defective chromodomain. In agreement with predictions from structural homology-based modeling, LHP1 with the mutated chromodomain had strongly reduced binding to H3K27me3 *in vitro*. Furthermore, recruitment to target genes and intra-nuclear localization of mutated LHP1 was greatly impaired *in vivo*. Because the phenotype of this new *lhp1* allele is very similar to an *lhp1* null allele, we conclude that chromodomain-mediated binding of LHP1 to H3K27me3 is essential for LHP1 function. These results support the model that LHP1 has a PRC1-like function in plants.

Results

An LHP1 mutant protein with a defective chromodomain

The new *lhp1-7* allele was discovered in a suppressor screen of a late flowering transgenic line with reduced MSI1 function (*msi1-tap1*; [18]). For details of the mutant screen see Materials and Methods. Sequencing of the *LHP1* locus and the *LHP1* cDNA

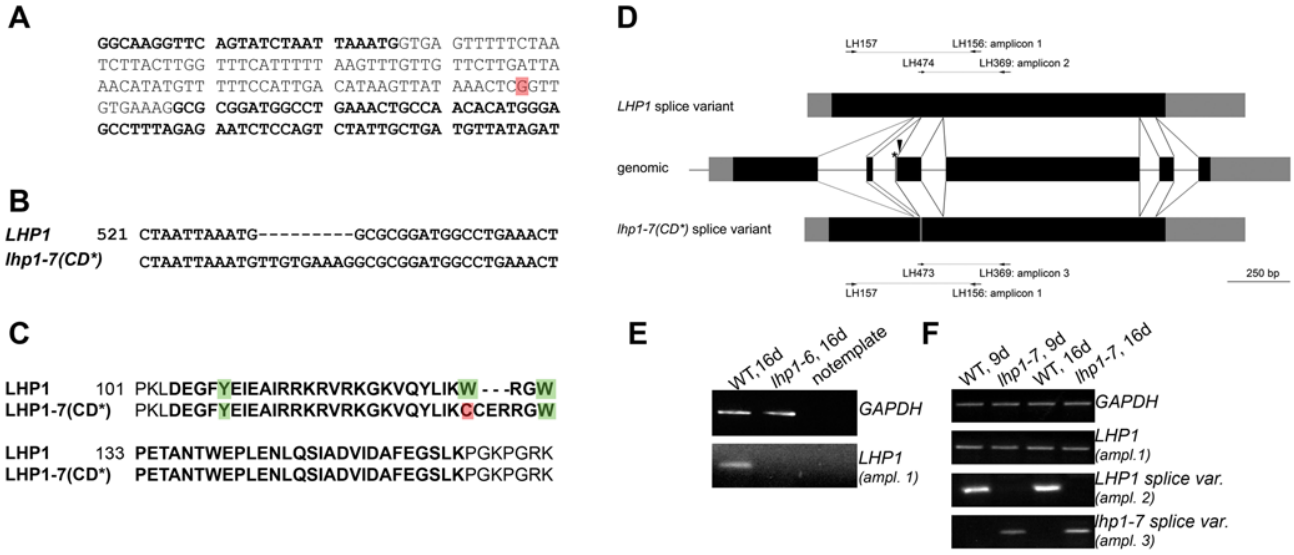


Figure 1. A novel *lhp1* allele. (A) Part of the genomic sequence of wild-type *LHP1*; exons are marked in bold. The EMS allele *lhp1-7* has a G to A transition in the second intron (position marked in red). (B) The point mutation in *lhp1-7* creates a new splice site. The critical region of the alignment of wild-type and mutant *LHP1* cDNAs is shown. (C) The lhp1-7 mutant protein has a defect in the chromodomain. The critical region of the alignment of the wild-type and mutant LHP1 proteins is shown. The chromodomain is shown in bold; the aromatic cage residues are marked in green and the cysteine that substitutes one of them in the mutant protein is marked in red. (D) The structure of wild-type and mutant *LHP1* transcripts and primers for PCR amplicons. Black boxes, grey boxes and lines represent exons, untranslated regions and introns, respectively. The asterisk marks the position of the point mutation in *lhp1-7* and the additional exon inclusion is shown in dark grey. The arrow marks the position of the T-DNA insertion in *lhp1-6*. Note that amplicon 1 is not specific for either splice variant while amplicons 2 and 3 are specific for the wild-type and mutant splice variants, respectively. For primer sequences see Table 3. (E) *lhp1-6* is a transcriptional null mutant. (F) *lhp1-7* expresses the mutant splice variant at the same level as the wild-type expresses *LHP1*. RNA in (E, F) was isolated from seedlings grown under long day conditions for 9d or 16d.
doi:10.1371/journal.pone.0005335.g001

revealed a newly created splice site that led to the presence of nine additional nucleotides at the junction of exons two and three in the processed *lhp1-7* transcript (Fig. 1A, B). This results in three additional amino acids (Cys-Glu-Arg) in the chromodomain adjacent to the conserved tryptophan 129, which is changed into a cysteine (Fig. 1C). The *lhp1-7* allele was introduced into the Columbia wild-type by backcrossing, and all further experiments were performed with *lhp1-7* in the wild-type background unless otherwise specified. We compared *lhp1-7* to the *lhp1-6* null allele, which we isolated previously from the SALK T-DNA insertion collection (line SALK_011762). While no *LHP1* transcript was detected in the *lhp1-6* T-DNA insertion mutant (Fig. 1D, E), *LHP1* transcript levels in *lhp1-7* were similar to those in wild-type (Fig. 1F). However, in *lhp1-7* only the mutant but not the wild-type splice variant was detected (Fig. 1F), suggesting that *lhp1-7* produces no or only very little wild-type protein. We discovered *lhp1-7* in a screen for suppressors of reduced MSI1 function, but the potential link between LHP1 and the histone binding WD40 repeat protein MSI1 will be discussed elsewhere (for a review about MSI1-like proteins see [19]). Here, we used the new *lhp1-7* allele to probe LHP1 function *in vitro* and *in vivo*.

The HP1 and Pc chromodomains have binding cavities formed by three aromatic residues to accommodate methylated lysines of H3 histone tails [3,20,21]. Homology-based modeling revealed that similar to HP1 and Pc, the chromodomain of LHP1 has the potential to form a binding cage containing three aromatic residues (Fig. 2A). Because one of the three aromatic residues, tryptophan 129, was changed to a cysteine in the chromodomain of lhp1-7, it is likely that this protein cannot form the typical binding cage and will be called LHP1-CD* (Fig. 2B). To more easily distinguish between the *lhp1-6* null allele and the *lhp1-7* mutant, we will refer to these alleles as *lhp1-6 (null)* and *lhp1-7 (CD*)*. Calculation of interaction energies suggested that LHP1-CD* has reduced affinity to trimethylated and unmethylated lysine residues (Table 1).

Next, we tested whether binding to H3K27me3 was indeed affected by the *lhp1-7 (CD*)* mutation. Similar to previously reported results, wild-type LHP1 bound strongly to the H3K27me3 peptide *in vitro*, but LHP1-CD* binding to H3K27me3 was significantly reduced and similar to the binding to unmethylated H3K27 (Fig. 2C). The reduced binding affinity to H3K27me3 *in vitro* suggests that LHP1-CD* could have compromised activity *in vivo*.

The LHP1 chromodomain is required for correct sub-nuclear localization and binding to target genes

To analyze the *in vivo* activity of LHP1-CD*, we introduced LHP1-GFP and LHP1-CD*-GFP fusion proteins into *lhp1-7(CD*)*. We found several lines in which the LHP1-GFP fusion protein could complement *lhp1-7(CD*)*, demonstrating that LHP1-GFP is fully functional (Fig. 3A, B). In contrast, the LHP1-CD*-GFP fusion protein was expressed (Fig. 3G, H) but unable to complement the mutant, suggesting that LHP1-CD*-GFP cannot substitute for wild-type LHP1.

Microscopic inspection of the LHP1-GFP and LHP1-CD*-GFP lines revealed that both wild-type and the mutant fusion proteins were targeted to the nucleus. The LHP1-GFP fusion protein showed a speckled pattern throughout the nucleus in most lines (Fig. 3C–F), similar to published data [13]. In contrast, the mutant LHP1-CD* was more uniformly distributed in the nucleus, often with additional strong accumulation in the nucleolus (Fig. 3G–K). Accumulation of mutant LHP1 versions in the nucleolus has been reported before [13,22], but the relevance of this abnormal targeting is unknown.

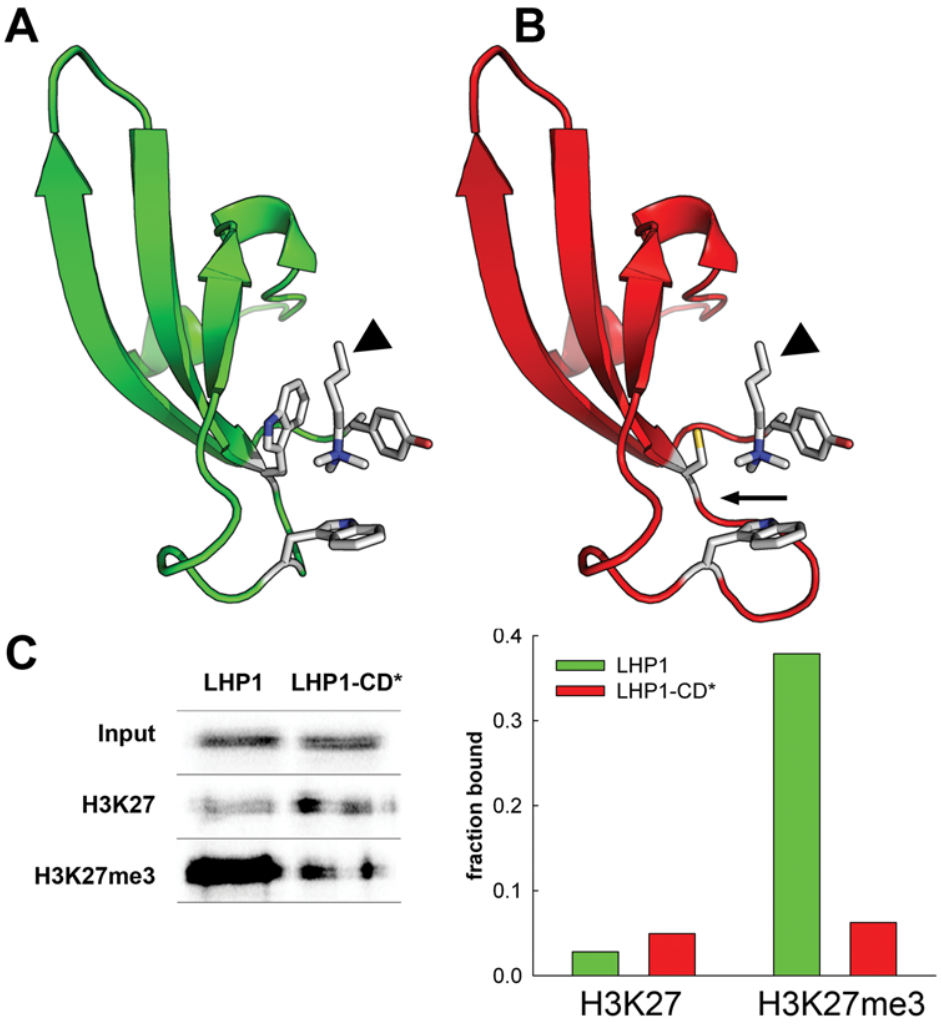


Figure 2. *lhp1-7(CD*)* encodes an LHP1 mutant protein with a defective chromodomain. (A) Structural model of the LHP1 chromodomain based on homology modeling using the coordinates of the Drosophila Pc chromodomain complexed with an H3K27me3 peptide [3]. (B) Structural model of the LHP1-CD* chromodomain, which is encoded by *lhp1-7(CD*)*. The arrow indicates the mutated region in LHP1-CD*. The position of the trimethylated lysine side chain (arrow heads) in (A) and (B) was derived from the template crystal structure. (C) Peptide-binding pull-down assay for wild-type LHP1 and LHP1-CD* (left) and quantification (right).
doi:10.1371/journal.pone.0005335.g002

Altered *in vitro* binding and sub-nuclear distribution of LHP1-CD* could also affect binding to individual target loci. We used the GFP fusion lines to test binding of LHP1 to *AGAMOUS* (*AG*) and *SEPALATA3* (*SEP3*), which are well-established PcG and LHP1 targets [14–16,23]. After chromatin immunoprecipitation we found that LHP1-GFP, but not LHP1-CD*-GFP, bound efficiently to both loci (Fig. 3L). Together, these results show that LHP1-CD* lost specificity for H3K27me3 *in vitro* and that LHP1-CD*-GFP cannot bind to at least some LHP1 targets *in vivo*, which may explain its altered sub-nuclear localization.

Table 1. Intermolecular energy values in Kcal/mol calculated by CHARMM upon minimization using a distance dependent dielectric function.

	Van der Waals	Electrostatic	Total
LHP1-CD/H3K27me3	−25.0	−3.9	−28.9
LHP1-CD*/H3K27me3	−22.5	−3.0	−25.5
LHP1-CD/H3K27	−19.5	−3.3	−22.8
LHP1-CD*/H3K27	−15.9	−2.5	−18.4

doi:10.1371/journal.pone.0005335.t001

Development is altered in *lhp1-7(CD*)* mutants

We compared the *lhp1-7(CD*)* mutant to wild-type and *lhp1-6*(null) mutant plants to establish which aspects of LHP1 function depend on chromodomain binding to H3K27me3. Analysis of flowering time revealed that both *lhp1-7(CD*)* and *lhp1-6(null)* plants flowered at similar times but much earlier than wild-type under long and short day conditions (Fig. 4). Early flowering was characterized by shortened juvenile and adult phases concomitant with strong *FT* upregulation (Fig. 4B, D). Epidermal cells of *lhp1* mutant rosette leaves were much smaller, although they maintained the characteristic jigsaw like shape (Fig. 5). Leaf cell number and expansion were reduced in both *lhp1* alleles, causing a strongly decreased rosette leaf size (Fig. 5).

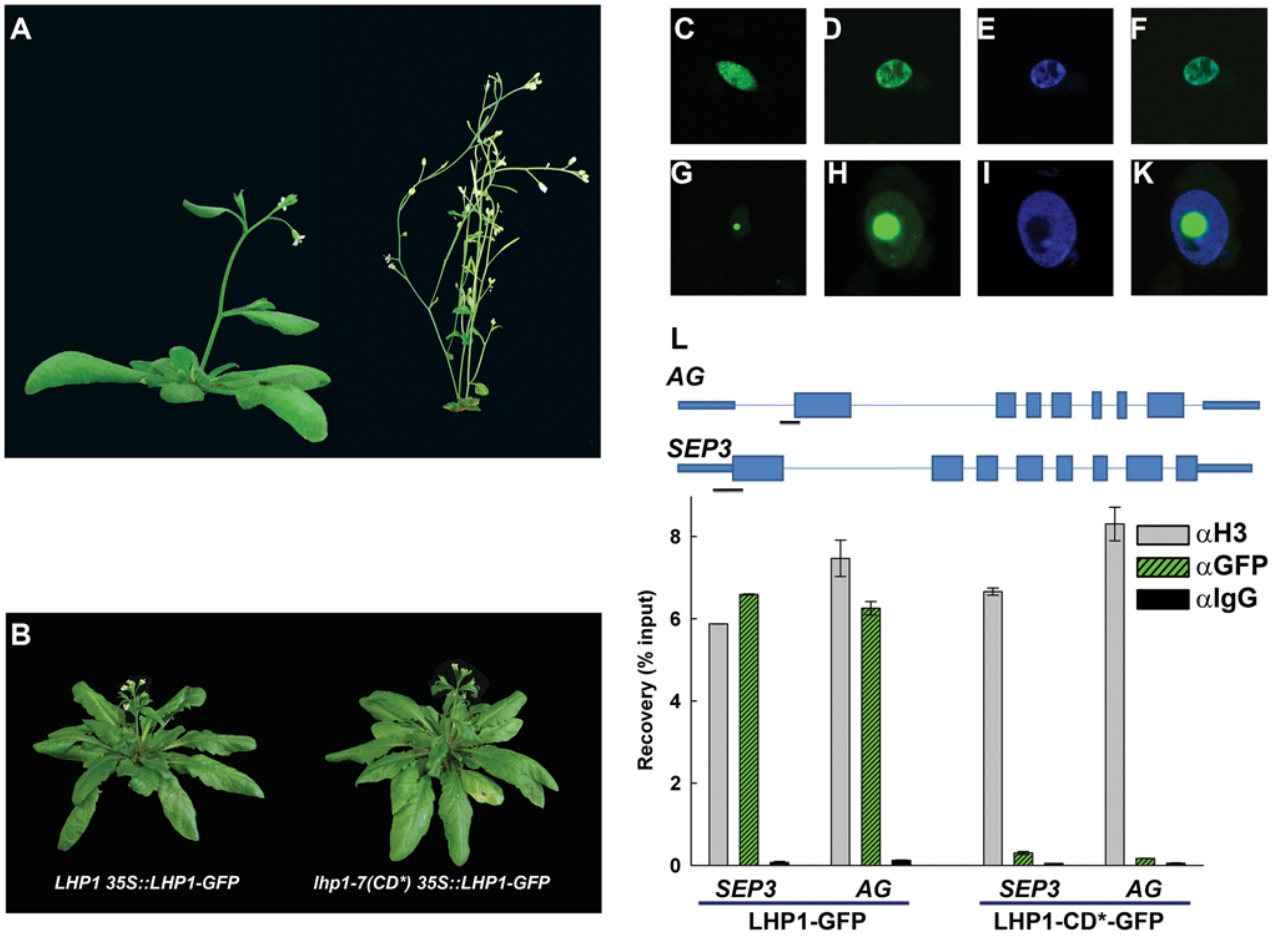


Figure 3. Altered sub-nuclear localization of LHP1-CD*-GFP. (A) Wild-type (Col, left) and *lhp1-7(CD*)* (right) after five weeks of growth under long day photoperiod. (B) *LHP1 35S::LHP1-GFP* (left) and *lhp1-7(CD*) 35S::LHP1-GFP* (right) plants. Plants are in the *msi1-tap1* background. (C-K) *35S::LHP1-GFP* (C-F) and *35S::lhp1-7(CD*)-GFP* plants (G-K) were used to analyze protein localization in leaf nuclei. Protein localization was detected by confocal laser scanning microscopy of GFP-fluorescence (C, G) or by immuno-localization (D, H). (E, I) DAPI-staining of the nuclei in D and H; merged images of D and E (F) and of H and I (K). (L) ChIP assays for binding of LHP1-GFP and LHP1-CD*-GFP to the *AG* and *SEP3* loci. Top: Genomic structure of *AG* and *SEP3*. Lines represent introns, narrow bars 3′ and 5′ UTRs and wide bars represent coding exons. Black lines represent regions probed by qPCR. Values are recovery as percent of input; IgG served as negative control.
doi:10.1371/journal.pone.0005335.g003

Arabidopsis LHP1 was initially identified genetically for its terminal flower phenotype [10]. Both *lhp1-6*(null) and *lhp1-7(CD*)* have the terminal flower phenotype, but *lhp1-7(CD*)* formed the terminal flower later than *lhp1-6(null)* (Fig. 6A). Consistently, primary stem growth ceased much earlier in *lhp1* mutants than in wild-type plants, but later in *lhp1-7(CD*)* than in *lhp1-6(null)* (Fig. 6B, C). In both *lhp1* alleles, not only duration of primary stem growth but also growth rates were reduced (Fig. 6D). Together, *lhp1-7(CD*)* is phenotypically similar to *lhp1-6(null)* during early plant development, but has a slightly milder phenotype late in development.

Silencing of PcG target genes is lost in *lhp1-7(CD*)* mutants

Flowers produced late during *lhp1-6* and *lhp1-7(CD*)* development often have supernumerary, missing or deformed organs (Fig. 7A–C), which may be caused by deregulation of floral homeotic genes. *AG* and *SEP3* were ectopically expressed in *lhp1-6(null)* and *lhp1-7(CD*)* rosette leaves (Fig. 7D). Similarly, *MEDEA* and *AGL19*, two PcG targets [24,25], were de-repressed in both *lhp1* alleles (Fig. 7D and data not shown). The observation that there was no reactivation of transposons or pseudogenes (*At4g03760*, *MU1*, *TA2*) or of targets of the RNA-dependent DNA-methylation pathway (*IG/LINE*, *IG2*, *IG5*, *RPL18*) (Fig. 7E and data not shown) confirmed that loss of LHP1 does not affect silencing in heterochromatin [13,14].

Together, our results show that similar to *lhp1-6(null)* major developmental regulatory genes (e.g., *FT*, *AG* and *SEP3*) are not repressed in *lhp1-7(CD*)* at times when they should be silent. Thus, we conclude that specific binding of LHP1 to H3K27me3 is essential to maintain repression of PcG target genes.

Discussion

In animals, PRC2 complexes set H3K27me3 marks, which assist to recruit PRC1 to mediate stable silencing [2]. Plant LHP1 proteins are similar to metazoan HP1, but could have PRC1 functions. Phylogenetic analysis suggests that the LHP1 and HP1 protein subfamilies have strongly diverged (Fig. 8). In addition to Arabidopsis, genes for LHP1 homologues were previously described for multiple mono- and dicotyledonous plant species

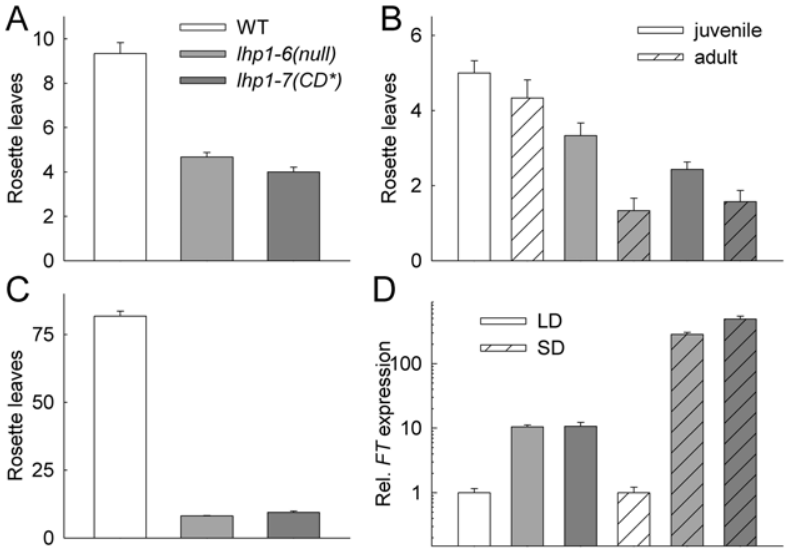


Figure 4. Early flowering of *lhp1* mutants. (A) Rosette leaves produced until bolting in long days (LD). (B) Phase transition in LD. (C) Rosette leaves formed until bolting in SD. Values in (A-C) are mean±S.E. (n≥7). (D) FT expression at ZT = 4h (ZT, *zeitgeber* time; ZT = 0 is lights on) in 12 days old seedlings from LD and at ZT = 6h in 14 days old seedlings from SD. Samples were taken at times when *FT* expression in wild-type is low [43]. Values in D are mean±S.E. (n = 4). Note that expression values for LD and SD were independently normalized to the corresponding wild-type. For all panels: White, grey, and dark-grey bars represent wild-type, *lhp1-6(null)* and *lhp1-7(CD*)*, respectively.
doi:10.1371/journal.pone.0005335.g004

such as apple, rape seed, carrot, tomato, rice and maize [11,12,26]. We found LHP1 homologues also in the genomes of poplar (*Populus trichocarpa)*, of a lycophyte (*Selaginella moellendorffii*), an ancient vascular plant lineage, and of a moss (*Physcomitrella patens*). In contrast, we failed to identify LHP1 or HP1 homologs in the genomes of the chlorophyte algae *Volvox carteri* and *Chlamydomonas reinhardtii*, suggesting that the presence of LHP1 is linked to multicellular development in the plant kingdom. Because chromatin immunoprecipitation has shown that LHP1 binding overlaps with H3K27me3 and LHP1 can bind H3K27me3 *in*

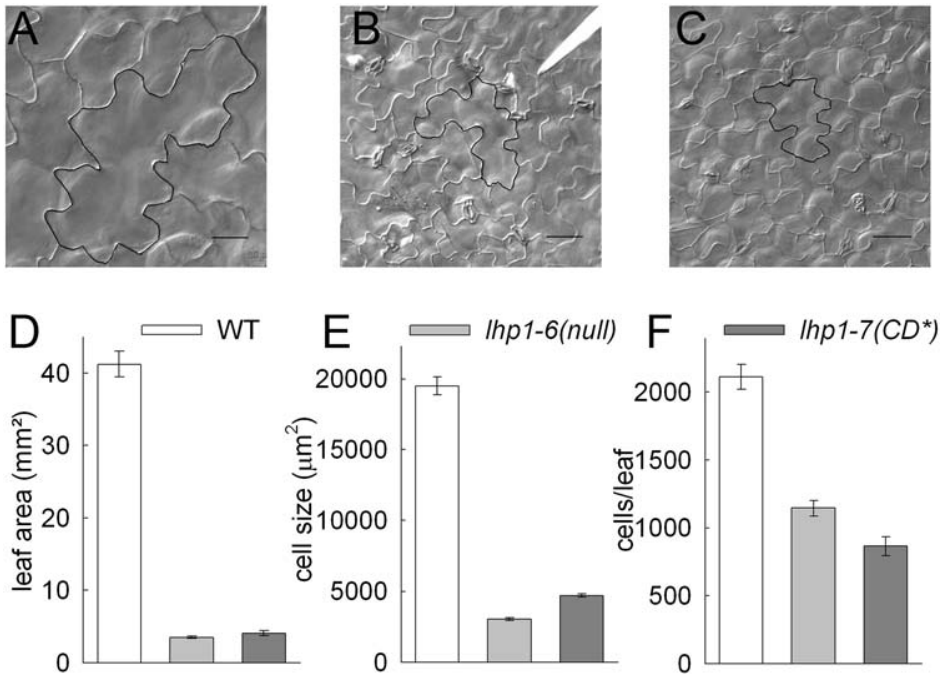


Figure 5. Altered leaf development in *lhp1* mutants. (A-C) Adaxial epidermis of Col (A), *lhp1-6(null)* (B) and *lhp1-7(CD*)* (C) leaves. (D) Area of first and second rosette leaf after bolting (n≥11). (E) Cell size in the adaxial epidermis of the first and second rosette leaves (n≥244). (F) Estimated cell number in the adaxial epidermis of the first and second rosette leaf. For all panels: White, grey, and dark-grey bars represent wild-type, *lhp1-6(null)* and *lhp1-7(CD*)*, respectively. Values in (D-F) are mean±S.E.
doi:10.1371/journal.pone.0005335.g005

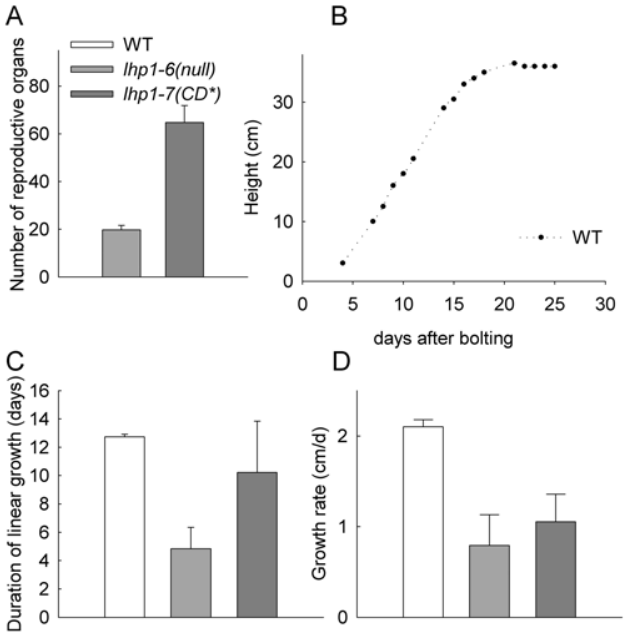


Figure 6. Altered shoot development in lhp1 mutants. (A) The total number of reproductive organs (siliques, flowers and flower buds) on the primary shoots of five weeks old plants from LD. Values are mean±S.E. (n≥8). (B) Example of the growth curve of a wild-type plant's primary shoot. (C) Length of linear growth phase. (D) Growth rates during linear growth phase of primary shoots. Values in (C, D) are averages over two experiments with n≥9 per experiment. For all panels: White, grey, and dark-grey bars represent wild-type, *lhp1-6(null)* and *lhp1-7(CD*)*, respectively.
doi:10.1371/journal.pone.0005335.g006

vitro, it was suggested that the chromodomain-protein LHP1 is a PRC1 equivalent of plants [15,16]. In contrast to animals, however, where PRC1 is needed for spreading of H3K27me3 over extended regions, in plants loss of LHP1 does not affect genomic H3K27me3 distribution [15].

Three aromatic residues form the binding cavity for methylated lysines of H3 in the chromodomain of animal HP1 and Pc [3,20,21,27]. Based on protein homology modeling, the chromodomain of plant LHP1 forms a similar binding pocket. Therefore we suggest that the novel lhp1 allele *lhp1-7(CD*)* has a defective binding pocket for the quaternary ammonium group because the preference of LHP1 for H3K27me3 over H3K27 was lost for LHP1-CD*. Energy calculations using CHARMM [28] and the CHARMm [29] force field are in qualitative agreement with the relative affinities measured by the pull-down assay. A quantitative agreement is not expected because of approximations inherent to the force field and the qualitative nature of the pull-down assays. An LHP1-CD*-GFP fusion did not efficiently bind to target gene chromatin and had lost its correct sub-nuclear distribution, suggesting that chromodomain-mediated binding to H3K27me3 is essential for LHP1 targeting *in vivo*. In contrast, the chromodomain might not be necessary for targeting of animal HP1 *in vivo* [30–32].

Mutations in Arabidopsis LHP1 strongly affect development [10,12]. The phenotype of the *lhp1-7(CD*)* allele was very similar to that of an *lhp1* null allele, suggesting that LHP1 function requires an intact chromodomain. Because only LHP1-GFP but not LHP1-CD*-GFP could rescue *lhp1* mutants, LHP1-CD* has no or strongly reduced biological activity. Residual binding of LHP1-CD* to H3K27me3 could explain the phenotypic differences between *lhp-7(CD*)* and *lhp1-6(null)* plants.

Loss of LHP1 or PRC2 share many similar developmental and molecular effects. Our experimental results, supported by homology modeling and previous reports, have revealed that LHP1 contributes to PRC1-like functions in plants and that chromodomain-mediated binding to H3K27me3 is required for this activity.

Materials and Methods

Plant material and growth conditions

All mutants used are in the Columbia (Col) wild-type accession of *Arabidopsis thaliana*. The *ddm1-2* allele was described before [33]. A new *lhp1* allele, *lhp1-6*, was identified in the SALK T-DNA

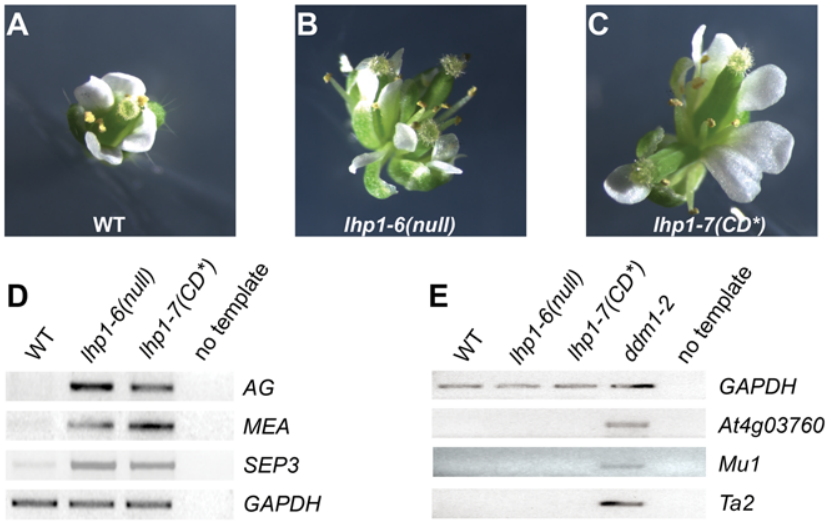


Figure 7. Loss of silencing at PcG targets and maintenance of silencing at heterochromatic loci in lhp1 mutants. (A–C) Flowers of wild-type Col (A), *lhp1-6(null)* (B) and *lhp1-7(CD*)* (C) produced late during development. (D) Expression of PcG targets in seedlings at ZT = 5h after 16 days in LD. (E) Expression of heterochromatic loci in rosette leaves at ZT = 5h after 25 days in LD. RNA from *ddm1-2* was used as positive control.
doi:10.1371/journal.pone.0005335.g007

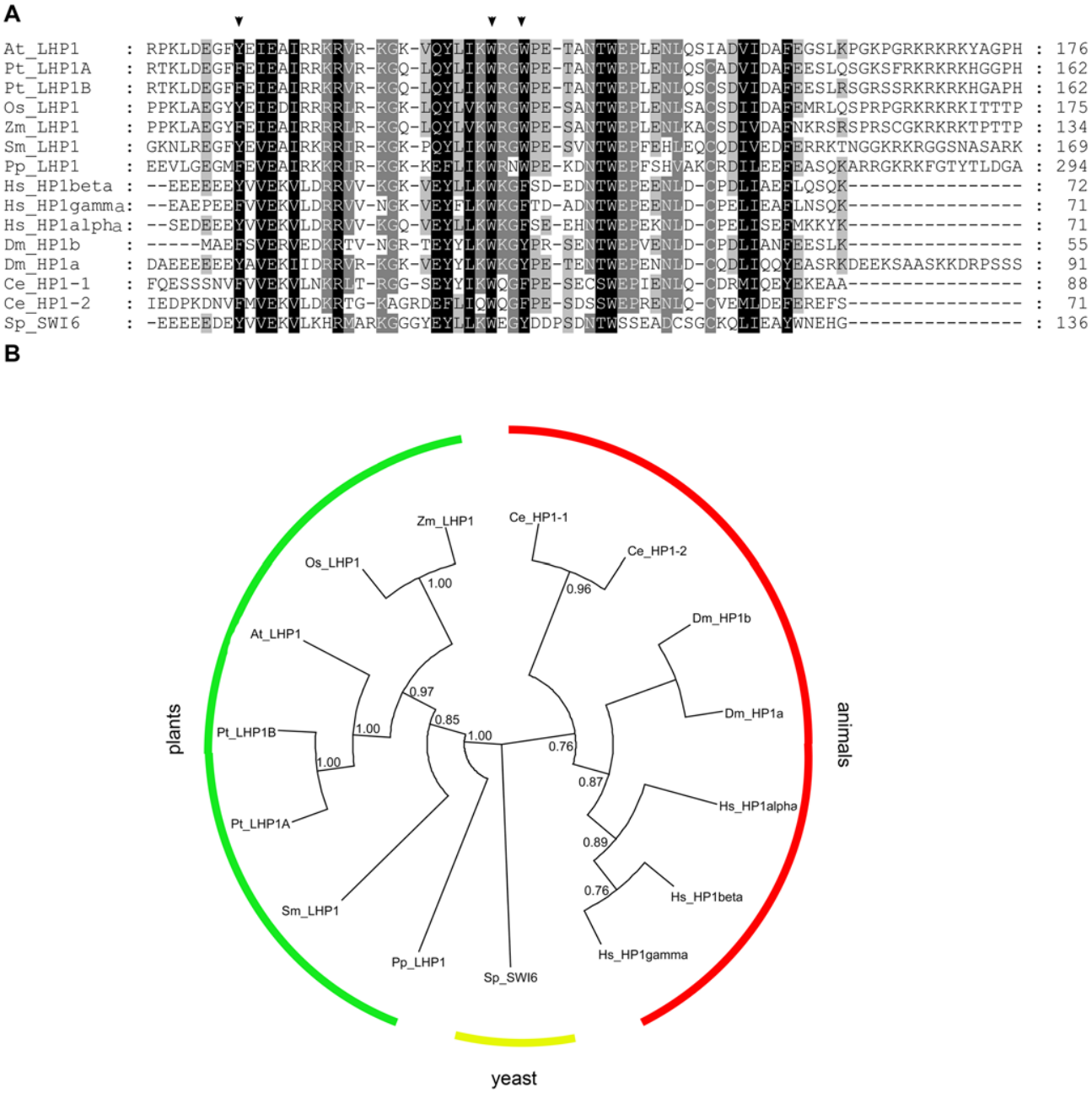


Figure 8. Sequence alignment and phylogenetic tree of LHP1 and HP1 homologues. (A) Segment of the alignment that contains the chromodomain. The arrow heads highlight the aromatic cage residues that form the binding cavity for histone methyl groups. (B) Phylogenetic tree of 15 LHP1 and HP1 homologs (*Arabidopsis thaliana*: At_LHP1 GI:15625407. *Populus trichocarpa*: Pt_LHP1A, estExt_Genewise1_v1.C_LG_XIX1329; Pt_LHP1B, eugene3.00130688. *Oryza sativa*: Os_LHP1, GI:110810411. *Zea mays*: Zm_LHP1 GI:22135459. *Physcomitrella patens*: Pp_LHP1, jgi|Phypa1_1|169812|estExt_fgenesh1_pg.C_2200058. *Selaginella moellendorffii*: Sm_LHP1, jgi|Selmo1|407083|fgenesh2_pg.C_scaffold_6000334. *Homo sapiens*: Hs_HP1alpha, GI:6912292; Hs_HP1beta, GI:48428808; Hs_HP1gamma, GI:5732187. *Drosophila melanogaster*: Dm_HP1a, GI:17136528; Dm_HP1b, GI:24640713. *Caenorhabditis elegans*: Ce_HP1-1, GI:17568757; Ce_HP1-2, GI:71987888. *Schizosaccharomyces pombe*: Sp_SWI6, GI:510930.). The evolutionary history was inferred using the Maximum Parsimony method; the most parsimonious tree with length = 2089 is shown. Support for each node, assessed with bootstrap analysis (1000 replicates) is given when higher than 60%. Note that the tree is displayed as circular cladogram with all branches of the same length.
doi:10.1371/journal.pone.0005335.g008

insertion mutant collection (SALK_011762). *LHP1* and *lhp1-7* cDNAs were cloned into vector pK7FWG2 [34], which was used to transform plants by floral dip with *Agrobacterium tumefaciens* (strain GV3101). Seeds were germinated on sterile basal salts Murashige and Skoog (MS) medium (Duchefa, Brussels, Belgium), and plants were analyzed on plates or transferred to soil 10 days after germination. Alternatively, seeds were directly sown on soil. Plants were kept in Conviron growth chambers with mixed cold

Table 2. Assays used for qPCR.

Gene	Forward primer	Reverse primer	Universal Probe Library probe
FT	GGTGGAGAAGACCTCAGGAA	GGTTGCTAGGACTTGGAACATC	#138 (Arabidopsis)
PP2A	GGAGAGTGACTTGGTTGAGCA	CATTCACCAGCTGAAAGTCG	#82 (Arabidopsis)
AG	CTAATCAAATTTTGCCCTAAACG	TCCTAGCTCCGATTGGTACG	#132 (Arabidopsis)
SEP3	ATTGATCTTGTTCTCTATCCTCTTCAA	AGAGAGAGAGATTGAGATATCTTTTGG	#103 (Arabidopsis)

doi:10.1371/journal.pone.0005335.t002

fluorescent and incandescent light (110 to 140 µmol m^{-2} s^{-1}, 21±2°C) under long day (LD, 16h light) or short day (SD, 8h light) photoperiods or were alternatively raised in green houses.

Isolation of the new *lhp1-7(CD*)* allele

For a suppressor screen, seeds of the late flowering *msi1-tap1* transgenic line [18] were mutagenised with ethyl methane sulfonate (EMS). Approximately one thousand F2 families were screened for suppression of the delayed floral transition of *msi1-tap1*. One family (0.3 362) segregated plants with a conspicuous early flowering phenotype. These early flowering plants were smaller, had reduced fertility and segregated in a 1:3 ratio (data not shown), suggesting recessive Mendelian inheritance. Molecular mapping located the mutation between the markers CER456657 (BAC MPI7) and CER457604 (BAC MXE10) on the top arm of chromosome V. Within the same region lies the gene *At5g17690*, which encodes LHP1. Because of similarities between the phenotypes of 0.3 362 plants and *lhp1* mutants, the *At5g17690* locus in 0.3 362 was sequenced and a single G to A transition was discovered.

To confirm that the mutation in the LHP1 gene is indeed responsible for the observed phenotype, an allelism test between 0.3 362 and the *lhp1-6* null allele was performed. The analyzed F1 and F2 generations displayed a homogenous appearance with small rosette size and were early flowering (data not shown), while genotyping revealed the expected ratios of plants homozygous, heterozygous or negative for the presence of the *lhp1-6* T-DNA insertion (data not shown), confirming that 0.3 362 was allelic to *lhp1-6*. The newly identified *lhp1* allele was henceforth called *lhp1-7*.

Flowering time and growth kinetics

Flowering time was defined as the time needed by the plants (n>7) to form a 5 mm high primary shoot. In addition, the numbers of juvenile and adult rosette leaves were determined based on the presence of abaxial trichomes as indicators for phase identity [35]. For growth kinetics, the height of the primary shoot was measured daily. The end of the linear growth phase was determined manually for individual plants from height vs. time after bolting diagrams. Primary shoots of wild-type plants grew linearly for nearly two weeks after bolting before growth ceased gradually (Fig. 6B).

In vitro transcription/translation and pull down assays

LHP1 and *lhp1-7* cDNAs were cloned into vector pRSET-A (Invitrogen) for *in vitro* transcription/translation reactions (TNT® T7 Quick Coupled Transcription/Translation System, Promega, Madison, WI) supplemented with L-[^{35}S]methionine. Equal amounts of wild-type and mutant protein were incubated with H3K27 or H3K27me3 peptides (LATKAARKSAPATGGC) coupled to SulfoLink Coupling Gel (Pierce Perbio, Lausanne, Switzerland). Samples were resolved by SDS-PAGE, exposed to a storage phosphor screen (Amersham Biosciences, Otelfingen, Switzerland) and visualized using a Molecular Imager FX Pro Plus System (BioRad, Reinach, Switzerland).

RNA isolation, RT-PCR and Real Time PCR

RNA isolation and RT-PCR was performed as previously described [36]. For Q-PCR analysis, the Universal Probe Library system (Roche Diagnostics, Rotkreuz, Switzerland) was used on a 7500 Fast Real-Time PCR instrument (Applied Biosystems, Lincoln, CA). *PP2A* was used as reference gene [37]. Q-PCR was performed with three to four replicates, and results were analyzed as described [38]. For details of the assays see Table 2.

Immuno-localisation

Immuno-localization of GFP fusion proteins was performed as described previously [39] using nuclei isolated from rosette and cauline leaves and a rabbit anti-GFP antibody (A11122, Molecular Probes Invitrogen, Basle, Switzerland). For detection, Alexa Fluor® 488 goat anti-rabbit IgG (A11008; Molecular Probes Invitrogen, Basle, Switzerland) was used. The preparations were analyzed by either epifluorescence microscopy (Zeiss Axioplan 2) or by confocal laser microscopy (Leica TCS SP1). For confocal laser microscopy, Alexa fluorophores were excited with a 488 nm laser; the emission signal was collected in a wavelength window between 502 nm and 543 nm. DAPI fluorescence was collected in a window from 438–485 nm

Structure determination by homology modeling

Wild-type and mutant sequences were processed by the SwissModel server [40] in automatic mode, fixing as a template the A chain of 1PDQ (Drosophila Polycomb chromodomain complexed with the histone H3 tail containing trimethyl lysine 27 [3]). CHARMm atom types and force field parameters [29] were assigned for all structures. Hydrogen atoms were added and minimized with the program CHARMM [28]. Trimethylated and unmethylated lysine residues were blocked with acetyl and N-methyl-aminyl groups. They were then minimized in the rigid protein using CHARMM and a distance-dependent dielectric

Table 3. Primer sequences.

Primer-ID	Sequence
LH156	TGCATATTTGCGCTTCCGTTT
LH157	CGGTGGAAACAGTCGGAGAAA
LH369	GGAAGGCTAGAGTTGTTGAGAGAC
LH473	GGTTCAGTATCTAATTAAATGTTGTGAAAG
LH474	GGCAAGGTTCAGTATCTAATTAAATGG

doi:10.1371/journal.pone.0005335.t003

function ($\varepsilon = 4r$). During minimization, harmonic constraints with a force constant of 2.5 Kcal/mol/Å2 were added to the blocking groups. The starting position of the trymethylated lysine residue was obtained by superimposing 1PDQ to each model. The interaction energy between the protein and the trymethylated/unmethylated lysine residue was calculated by INTE command of CHARMM. Given the approximations inherent to the force field and the homology models, only a qualitative agreement with experimental data is expected.

Chromatin Immunoprecipitation

Chromatin isolation was performed as described previously [24] using 15d-old seedlings. Chromatin immunoprecipitation was done using the LowCell# ChIP kit (Diagenode, Liège, Belgium) according to manufacturer's instructions. The following antibodies were used: Polyclonal anti-H3 antibody (#01-690, Upstate, Charlottesville, VA), polyclonal anti-GFP antibody (#A11122, Molecular Probes Invitrogen, Basle, Switzerland) and non-immun IgG (Diagenode). Presence of *AG* and *SEP3* fragments was determined by qPCR using the Universal Probe system (Roche).

Sequence alignment and phylogenetic analysis

Protein sequences of HP1 and LHP1 proteins were selected based on previous publications [16,26] and on BLAST searches with the Arabidopsis LHP1 sequence using the DOE Joint Genome Institute data base (http://genome.jgi-psf.org/). Final sequence alignments of the selected sequences were generated with CLUSTALX 1.81 (protein weight matrix was Gonnet 250, gap opening penalty was 10.0, and gap extension penalty was 0.2). The evolutionary history was inferred using the flat-weighted Maximum Parsimony method. The MP tree of amino acid sequences was obtained using the Close-Neighbor-Interchange algorithm with search level 3 in which the initial trees were obtained with the random addition of sequences (10 replicates). All alignment gaps were treated as missing data. There were a total of 985 positions in the final dataset, out of which 292 were parsimony informative. Phylogenetic analyses were conducted in MEGA4 [41]. The presentation of the phylogenic tree was prepared using Dendroscope [42].

Acknowledgments

We thank Tobias Ludwig for help with molecular cloning of the expression constructs and the Light Microscopy Centre of the ETH Zurich for providing access to their infrastructure.

Author Contributions

Conceived and designed the experiments: VE CK LH. Performed the experiments: VE EA HS TW PA LH. Analyzed the data: VE EA HS PA AC CK. Wrote the paper: VE PA AC WG CK LH.

References

1. Köhler C, Villar CB (2008) Programming of gene expression by Polycomb group proteins. Trends Cell Biol 18: 236–243.
2. Schwartz YB, Pirrotta V (2007) Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet 8: 9–22.
3. Fischle W, Wang Y, Jacobs SA, Kim Y, Allis CD, et al. (2003) Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev 17: 1870–1881.
4. Aubert D, Chen L, Moon YH, Martin D, Castle LA, et al. (2001) EMF1, a novel protein involved in the control of shoot architecture and flowering in Arabidopsis. Plant Cell 13: 1865–1875.
5. Calonje M, Sanchez R, Chen L, Sung ZR (2008) EMBRYONIC FLOWER1 participates in Polycomb group-mediated *AG* gene silencing in Arabidopsis. Plant Cell 20: 277–291.
6. Mylne JS, Barrett L, Tessadori F, Mesnage S, Johnson L, et al. (2006) LHP1, the Arabidopsis homologue of HETEROCHROMATIN PROTEIN1, is required for epigenetic silencing of *FLC*. Proc Natl Acad Sci U S A 103: 5012–5017.
7. Sanchez-Pulido L, Devos D, Sung ZR, Calonje M (2008) RAWUL: A new Ubiquitin-like domain in PRC1 Ring finger proteins that unveils putative plant and worm PRC1 orthologs. BMC Genomics 9: 308.
8. Haughn GW, Davin L, Giblin M, Underhill EW (1991) Biochemical Genetics of Plant Secondary Metabolites in *Arabidopsis thaliana*: The Glucosinolates. Plant Physiol 97: 217–226.
9. Kim JH, Durrett TP, Last RL, Jander G (2004) Characterization of the Arabidopsis *TU8* glucosinolate mutation, an allele of *TERMINAL FLOWER2*. Plant Mol Biol 54: 671–682.
10. Larsson AS, Landberg K, Meeks-Wagner DR (1998) The *TERMINAL FLOWER 2* (*TFL2*) gene controls the reproductive transition and meristem identity in *Arabidopsis thaliana*. Genetics 149: 597–605.
11. Kotake T, Takada S, Nakahigashi K, Ohto M, Goto K (2003) Arabidopsis *TERMINAL FLOWER 2* gene encodes a LIKE HETEROCHROMATIN PROTEIN 1 homolog and represses both *FLOWERING LOCUS T* to regulate flowering time and several floral homeotic genes. Plant Cell Physiol 44: 555–564.
12. Gaudin V, Libault M, Pouteau S, Juul T, Zhao G, et al. (2001) Mutations in *LIKE HETEROCHROMATIN PROTEIN 1* affect flowering time and plant architecture in Arabidopsis. Development 128: 4847–4858.
13. Libault M, Tessadori F, Germann S, Snijder B, Fransz P, et al. (2005) The Arabidopsis LHP1 protein is a component of euchromatin. Planta 222: 910–925.
14. Nakahigashi K, Jasencakova Z, Schubert I, Goto K (2005) The Arabidopsis HETROCHROMATIN PROTEIN 1 homolog (TERMINAL FLOWER2) silences genes within the euchromatic region but not genes positioned in heterochromatin. Plant Cell Physiol 46: 1747–1756.
15. Turck F, Roudier F, Farrona S, Martin-Magniette ML, Guillaume E, et al. (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of Histone H3 Lysine 27. PLoS Genet 3: 0855–0866.
16. Zhang X, Germann S, Blus BJ, Khorasanizadeh S, Gaudin V, et al. (2007) The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. Nat Struct Mol Biol 14: 869–871.
17. Xu L, Shen WH (2008) Polycomb silencing of *KNOX* genes confines shoot stem cell niches in Arabidopsis. Curr Biol 18: 1966–1971.
18. Bouveret R, Schönrock N, Gruissem W, Hennig L (2006) Regulation of flowering time by Arabidopsis MSI1. Development 133: 1693–1702.
19. Hennig L, Bouveret R, Gruissem W (2005) MSI1-like proteins: an escort service for chromatin assembly and remodeling complexes. Trends Cell. Biol. 15: 295–302.
20. Jacobs SA, Khorasanizadeh S (2002) Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. Science 295: 2080–2083.
21. Nielsen PR, Nietlispach D, Mott HR, Callaghan J, et al. (2002) Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature 416: 103–107.
22. Zemach A, Li Y, Ben-Meir H, Oliva M, Mosquna A, et al. (2006) Different domains control the localization and mobility of LIKE HETEROCHROMATIN PROTEIN1 in Arabidopsis nuclei. Plant Cell 18: 133–145.
23. Germann S, Juul-Jensen T, Letarnec B, Gaudin V (2006) DamID, a new tool for studying plant chromatin profiling in vivo, and its use to identify putative LHP1 target loci. Plant J 48: 153–163.
24. Schönrock N, Bouveret R, Leroy O, Borghi L, Köhler C, et al. (2006) Polycomb-group proteins repress the floral activator *AGL19* in the *FLC*-independent vernalization pathway. Genes Dev. 20: 1667–1678.
25. Katz A, Oliva M, Mosquna A, Hakim O, Ohad N (2004) FIE and CURLY LEAF Polycomb proteins interact in the regulation of homeobox gene expression during sporophyte development. Plant J 37: 707–719.
26. Mimida N, Kidou SI, Kotoda N (2007) Constitutive expression of two apple (*Malus*×*domestica* Borkh.) homolog genes of LIKE HETEROCHROMATIN PROTEIN1 affects flowering time and whole-plant growth in transgenic Arabidopsis. Mol Genet Genomics 278: 295–305.
27. Min J, Zhang Y, Xu RM (2003) Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. Genes Dev 17: 1823–1828.
28. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, et al. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem 4: 187–217.
29. Momany FA, Rone R (1992) Validation of the general purpose QUANTA®3.2/CHARMm® force field. J Comp Chem 13: 888–900.
30. Cowell IG, Aucott R, Mahadevaiah SK, Burgoyne PS, Huskisson N, et al. (2002) Heterochromatin, HP1 and methylation at lysine 9 of histone H3 in animals. Chromosoma 111: 22–36.
31. Meehan RR, Kao CF, Pennings S (2003) HP1 binding to native chromatin *in vitro* is determined by the hinge region and not by the chromodomain. EMBO J 22: 3164–3174.
32. Dialynas GK, Makatsori D, Kourmouli N, Theodoropoulos PA, McLean K, et al. (2006) Methylation-independent binding to histone H3 and cell cycle-dependent incorporation of HP1beta into heterochromatin. J Biol Chem 281: 14350–14360.
33. Vongs A, Kakutani T, Martienssen RA, Richards EJ (1993) *Arabidopsis thaliana* DNA methylation mutants. Science 260: 1926–1928.

34. Karimi M, Inze D, Depicker A (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci 7: 193–195.
35. Telfer A, Bollman KM, Poethig RS (1997) Phase change and the regulation of trichome distribution in *Arabidopsis thaliana*. Development 124: 645–654.
36. Hennig L, Taranto P, Walser M, Schönrock N, Gruissem W (2003) Arabidopsis MSI1 is required for epigenetic maintenance of reproductive development. Development 130: 2555–2565.
37. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible WR (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. Plant Physiol 139: 5–17.
38. Simon P (2003) Q-Gene: processing quantitative real-time RT-PCR data. Bioinformatics 19: 1439–1440.
39. Schönrock N, Exner V, Probst A, Gruissem W, Hennig L (2006) Functional genomic analysis of CAF-1 mutants in *Arabidopsis thaliana*. J. Biol. Chem. 281: 9560–9568.
40. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31: 3381–3385.
41. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
42. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.
43. Suarez-Lopez P, Wheatley K, Robson F, Onouchi H, Valverde F, et al. (2001) CONSTANS mediates between the circadian clock and the control of flowering in Arabidopsis. Nature 410: 1116–1120.

Chapter 5

A hypersensitive CRYPTOCHROME 1 allele of Arabidopsis promotes flowering

Vivien Exner, Cristina Alexandre, Gesa Rosenfeldt, Pietro Alfarano, Mena Nater, Amedeo Caflisch, Wilhelm Gruissem, Alfred Batschauer, and Lars Hennig.

To be submitted.

May 26, 2010

A hypersensitive CRYPTOCHROME 1 allele of Arabidopsis promotes flowering

Vivien Exner[1], Cristina Alexandre[1], Gesa Rosenfeldt[2], Pietro Alfarano[3], Mena Nater[1], Amedeo Caflisch[3], Wilhelm Gruissem[1], Alfred Batschauer[2], Lars Hennig[1]

[1] Department of Biology & Zurich-Basel Plant Science Center, ETH Zurich, Universitätstrasse 2, 8129 Zurich, Switzerland

[2] FB Biologie/Pflanzenphysiologie, Philipps-Universität, Karl-von-Frisch-Str. 8. D-35032 Marburg, Germany

[3] Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Running title: A hypersensitive CRYPTOCHROME 1 allele

Corresponding author: Hennig, L. (lhennig@ipw.biol.ethz.ch)

Financial source:

This work was supported by SNF grant 3100AO-116060 (to L.H.), ETH project TH-16/05-2 (to L.H.), and DFG grant BA985/11-1 (to A.B.).

SUMMARY

Plants use different classes of photoreceptors to collect information about their light environment. Cryptochromes are blue light photoreceptors that control deetiolation, entrain the circadian clock and are involved in flowering time control. Here, we describe the *cry1-L407F* allele, which encodes a hypersensitive cryptochrome 1 protein. Plants carrying the *cry1-L407F* point mutation have elevated expression of *CO* and *FT* under short day conditions leading to very early flowering. These results demonstrate that not only the well-studied cryptochrome 2 with an unequivocal role in flowering promotion but also cryptochrome 1 can function as an activator of the floral transition. The *cry1-L407F* mutants are also hypersensitive towards blue, red and far-red light in hypocotyl growth inhibition. In addition, *cry1-L407F* seeds are hypersensitive to germination-inducing red-light pulses, but the far-red reversibility of this response is not compromised. This demonstrates that the cry1-L407F photoreceptor can increase the sensitivity of phytochrome signaling cascades. Molecular dynamics simulation of wild-type and the mutant cry1 proteins indicated that the L407F mutation considerably reduces the structural flexibility of two solvent-exposed regions of the protein, suggesting that the hypersensitivity might result from a reduced entropy penalty of binding events during downstream signal transduction.

Key-words: Arabidopsis, CO, cryptochrome, FT, flowering time, phytochrome

INTRODUCTION

Light determines the plant's life, because light is the essential energy source for plant metabolism. The spatial, temporal and spectral variability of light provides cues about the time of day, the season and the presence of competitors for light. Sensitive and precise light perception is therefore essential to properly adjust plant development for maximal photosynthetic efficiency, to correlate vegetative and reproductive growth with favorable seasons and eventually to maximize fitness. To cope with this task plants have evolved several types of photoreceptors including the phytochromes and cryptochromes (for review see (Banerjee et al., 2005, Josse et al., 2008, Muller et al., 2009). Phytochromes are red and far-red light receptors and regulate different aspects of plant development, such as hypocotyl elongation in red and far-red light and shade avoidance responses (Franklin et al., 2005). In addition, the two major phytochromes in Arabidopsis, phytochrome (phy) A and phyB are involved in flowering time control: phyA promotes flowering under short day (SD) and long day (LD) photoperiods (Johnson et al., 1994), while phyB acts as a floral inhibitor (Reed et al., 1993, Mockler et al., 1999).

Cryptochromes (cry) are flavoproteins with two chromophores that sense blue and ultraviolet-A light in plants (Lin et al., 2005). The essential chromophore is a flavin adenine dinucleotide (FAD), the second chromophore is supposed to function in light harvesting and is a pterine (methenyltetrahydrofolate) (Muller et al., 2009). Cryptochromes have an amino-terminal photolyase related (PHR) domain that is similar to photolyases, but are distinguished from the latter by a variable carboxy-terminal domain. Furthermore, crystallization of the photolyase-like domain of Arabidopsis cry1 has revealed additional structural differences between photolyase and cryptochrome that explain the lack of DNA-binding activity of the cryptochromes

(Brautigam et al., 2004). Moreover, the crystal structure confirmed the previously described ATP binding of cry1 (Bouly et al., 2003) by soaking cry1 crystals with the nonhydrolyzable ATP analog AMP-PNP and showing AMP-PNP located in the FAD access cavity (Brautigam et al., 2004). Despite the available protein structure, cryptochrome's mode of action still remains to be determined. Interestingly, the carboxy terminus of plant cryptochromes as well as a carboxy terminal 80-residue motif (NC80) confer constitutive cryptochrome signaling when overexpressed in plants even in the absence of light (Yang et al., 2000, Yu et al., 2007). For plant cryptochromes, it has been proposed that the PHR domain and the carboxy terminus form a "closed" conformation to mask the NC80 motif in the absence of light. Blue light would trigger phosphorylation of the carboxy terminal tail and its electrostatical repulsion from the surface of the PHR domain to form an "open" conformation, exposing the NC80 motif and initiating signal transduction (Yu et al., 2007).

The *Arabidopsis thaliana* genome harbors two cryptochrome genes: *CRY1* and *CRY2*. A third member of this family, *CRY3*, belongs to the DASH-type subgroup with repair activity for cyclobutane pyrimidine dimer lesions in single-stranded DNA (Selby et al., 2006) and loop structures of duplex DNA (Pokorny et al., 2008), but with so far unproven photoreceptor function. Mutations in both *CRY1* and *CRY2* interfere with inhibition of hypocotyl elongation under blue light conditions (Ahmad et al., 1993, Guo et al., 1998). Current data suggest that cry1 is the major blue light receptor for seedling photomorphogenesis, while cry2 is more important for the control of flowering time (Guo et al., 1998, Mockler et al., 1999, El-Din El-Assal et al., 2003, Mockler et al., 2003, Endo et al., 2007). Nevertheless, several studies also reported cry1 as a floral regulator: some *cry1* mutant alleles conferred late flowering under certain growth conditions (Bagnall et al., 1996, Blazquez et al., 2003), but others did not (Zagotta et al., 1996, Mockler et al., 1999).

Even though the functions of the different photoreceptors are assigned to specific segments of the spectrum of light, physiological and mutant analyses have revealed extensive cross-talk between blue and red light photoreceptors (Casal 2000): phyA and phyB display antagonistic and synergistic effects on the action of each other, depending on which responses are studied (Reed et al., 1994, Casal et al., 1995), and several of the phytochromes functionally interact with the cryptochromes (Ahmad et al., 1997, Casal et al., 1998, Neff et al., 1998, Hennig et al., 1999, Mockler et al., 1999, Devlin et al., 2000). In addition, a physical interaction of phyA with cry1 (Ahmad et al., 1998) and of phyB with cry2 (Mas et al., 2000) has been reported. While these cross-talks are of minor importance under controlled monochromatic light conditions, they are probably essential for fine tuning of developmental programs in natural environments.

Here, we provide evidence that a hyperactive *cry1* allele confers hypersensitivity not only to blue but also to red light and strongly shortens flowering time under SD.

RESULTS

Isolation of a new *cry1* allele

For a suppressor screen, seeds of the late flowering *msi1-tap1* transgenic line (Bouveret et al., 2006) were mutagenised with ethyl methane sulfonate (EMS) (Exner et al., 2009). Flowering time was scored under LD photoperiods for 1045 M_2 families. Among these, eleven families segregated plants that shortened the vegetative phase of *msi1-tap1*. For six of them we confirmed the phenotype in subsequent generations. One of these six contained a mutation in *LHP1* (Exner et al., 2009); the others belong to at least 4 complementation groups (data not shown). All these mutants still react to

changes in day length, but exhibited different responses as tested by flowering time experiments under LD and SD (data not shown). The mutant with ID 0.3-457 was chosen for further characterization (Tab. 1).

The mutation in 0.3-457 was localized between markers CER446440 (BAC T3H13) and CER460528 (BAC T3H13) on the lower arm of chromosome IV. This region contains 9 genes including *CRY1*. Sequencing of the *CRY1* locus revealed a C to T transition in the third exon, 1469 bp after the ATG start codon (Fig. 1A), which caused a change of leucine 407 into a phenylalanine (Fig. 1B). Therefore, 0.3-457 is henceforth called *cry1-L407F*.

Comparison of cryptochrome and photolyase sequences from different organisms revealed that leucine 407, which is located in the photolyase-like PHR domain, is conserved among nearly all plant cryptochromes. The only exception is Arabidopsis CRY2, which has an isoleucine instead of the leucine at this position (Fig. 1B). In contrast, photolyases and animal cryptochromes have amino acids with small side chains (alanine, serine, threonine or glycine) at this position but never a bulky hydrophobic residue such as leucine. Furthermore, leucine 407 is located in a block of 12 amino acids that are highly conserved in plant cryptochromes but not in photolyases and animal cryptochromes. A structure of the cry1 PHR domain has been reported (Brautigam et al., 2004), and according to this structure leucine 407 is buried inside the protein without solvent contact (Fig. 1C).

cry1-L407F strongly accelerates flowering

The *cry1-L407F* mutant did not only flower earlier in the *msi1-tap1* background but also when backcrossed into Col wild type both under long day (LD) and short day (SD) conditions (Fig. 2A, B). Under LD, the acceleration of flowering in *cry1-L407F* was caused by a shortened adult phase (0.8 vs. 2.6 leaves) while the

duration of the juvenile phase (5.3 leaves) was not affected. The early flowering phenotype was even more dramatic under SD. While wild-type plants produced 61 leaves before bolting, *cry1-L407F* produced only 12 leaves. Tests for genetic interaction between *cry1-L407F* and *msi1-tap1* revealed additivity of both mutants (Supplemental Fig. S1), suggesting that *CRY1* and *MSI1* function in separate genetic pathways in the control of flowering. Besides the early flowering phenotype, *cry1-L407F* featured a small and compact rosette (Supplemental Fig. S1), which is partly caused by shortened petioles (data not shown). Such a reduction of petiole length has also been described for plants overexpressing cry1 (Lin et al., 1996).

These results, together with the observation that *cry1-L407F* behaved semidominantly (data not shown), suggested that we identified a gain-of-function *CRY1* allele. Because genetic complementation tests and transgenic complementation are difficult with hyper- and neomorphic alleles, the ability of cry1-L407F to accelerate flowering was tested by transgenic phenocopy experiments. When wild-type cry1 was introduced and overexpressed in Col wild-type plants, the four tested transgenic lines all flowered as late or even slightly later than Col wild type under SD (Fig. 2C). Similarly, Lin et al. had previously reported that increased cry1 dosage caused slightly delayed flowering (Lin et al., 1996). In contrast, when the *cry1-L407F* mutant gene under control of the same CaMV 35S promoter was introduced into Col, the four tested transgenic lines all flowered earlier than Col wild type under SD (Fig. 2C).

Together, the results show that *cry1-L407F* is a gain-of-function *CRY1* allele that strongly accelerates flowering.

cry1-L407F causes *FT* over-expression

Cryptochromes can affect flowering in at least three different ways. First, cryptochromes control the phase of the circadian clock, which in turn controls diurnal expression patterns of *CONSTANS* (*CO*). Second, cryptochromes stabilize CO protein in the light, and CO then activates flowering by inducing expression of *FLOWERING LOCUS T (FT)*, which encodes the mobile flowering signal FT (for review see (Kobayashi et al., 2007). Third, cry2 can directly induce *FT* expression (Liu et al., 2008). To elucidate whether one of these mechanisms is involved in accelerated flowering of *cry1-L407F*, we measured gene expression of *ELF4*, *GI*, *CO* and *FT* under SD conditions. ELF4 and GI participate in signaling from the circadian clock to downstream processes such as *CO* expression (Park et al., 1999, Doyle et al., 2002). In SD, *CO* is usually expressed only after the end of the light phase, and because CO protein is rapidly degraded in the dark, *FT* remains inactive. We found that *ELF4* expression was not significantly affected in *cry1-L407F* maintaining its typical evening peak (Fig. 3), suggesting that the accelerated flowering was not caused by a malfunction of the circadian clock. Likewise, GI expression was very similar between wild type and *cry1-L407F* (Fig. 3). In contrast, expression of *CO* and *FT* was considerably increased in *cry1-L407F* (Fig. 3). However, *FT* expression was strongest during the light period (*zeitgeber* time (ZT) = 3h) while *CO* expression was strongest during the dark period (ZT = 14h), suggesting that cry1-L407F does not establish aberrant CO protein stabilization in the dark.

Together, these results indicate that cry1-L407F causes untimely expression of *CO* during the light period and thus allows for the accumulation of CO under SD photoperiods. Increased CO levels then strongly activate *FT* and cause the very early flowering of wild-type plants and the suppression of *msi1-tap1*'s late flowering phenotype.

cry1-L407F causes hypersensitivity towards blue and red light

The cry1 photoreceptor is known to control hypocotyl growth in response to blue light (Koornneef et al., 1980, Ahmad et al., 1993). To investigate the effects of the amino acid substitution on the function of cry1-L407F in further detail, seedlings were grown under different fluence rates of blue, red and far-red light and hypocotyl elongation was measured. Under blue light, inhibition of hypocotyl elongation was observed under much lower fluence rates in *cry1-L407F* than in wild type (Fig. 4). Thus, cry1-L407F is a hypersensitive blue-light photoreceptor, and *cry1-L407F* is a hypermorphic allele. The hypersensitivity of *cry1-L407F* towards blue light is a dominant trait. Heterozygous *cry1-L407F* seedlings uniformly displayed short hypocotyls when grown under low fluence rates of blue light (data not shown). Surprisingly, *cry1-L407F* seedlings were not only strongly hypersensitive to blue but also to red light (Fig. 4), which should not efficiently activate cry1 (Lin et al., 1995, Ahmad et al., 2002). It is unlikely that this phenotype was caused by a "contamination" of the red light by photons from the blue part of the spectrum, because there is extremely little if any blue light emitted the light source used in these experiments (Supplemental Fig. S2). Similar to the situation under red light, *cry1-L407F* seedlings were also hypersensitive to far red light, which is believed to be predominately sensed by phyA.

Repression of hypocotyl elongation under blue light is a normal function of wild-type Arabidopsis cry1 and cry2 (Koornneef et al., 1980, Ahmad et al., 1993, Lin et al., 1998). In order to test whether cry1-L407F does also affect red-light sensitivity of a process normally not controlled by cryptochromes, we compared light-dependent germination of wild type and *cry1-L407F*. It is commonly thought that germination of Arabidopsis seeds is exclusively controlled by phytochromes but not by

cryptochromes (Shinomura et al., 1996, Oh et al., 2007). Under continuous white light, both wild type and *cry1-L407F* show a similarly high frequency of germination, while in the dark or after a far red pulse almost no germination occurred (Fig. 5). Pulses of red or white light that caused a germination rate of about 30% in wild-type seeds caused a germination rate of 80% in *cry1-L407F* seeds. Thus, cry1-L407F strongly increased the sensitivity of red-light induced germination. Red light-induced germination is usually a function of phyB. However, the results shown in Figure 5 could indicate that *cry1-L407F* itself can induce germination. Because photoreversibility is a hallmark of phyB function, we tested whether the red light-induced induction of germination could be reverted by a pulse of far red light. Indeed, a far red light pulse could completely prevent red light-induced germination in both wild type and *cry1-L407F* (Fig. 5). This strongly suggests that cry1-L407F can increase the sensitivity to red light or signaling of photoreversible phytochromes such as phyB in the low fluence repsonse.

Phytochromes regulate transcription of many genes; *EARLY LIGHT INDUCED 2* (*ELIP2*), for instance, is rapidly upregulated after exposure to red or far red light, and this upregulation requires phyB and phyA, respectively (Harari-Steinberg et al., 2001). We tested whether the *cry1-L407F* mutation would affect the red light-induced expression of *ELIP2*. In darkness and during the investigated time course up to six hours after transfer to red light, ELIP2 expression was consistently higher in *cry1-L407F* than in wild type, but with very similar kinetics for both genotypes (Fig. 6). Thus, the effect of cry1-L407F on red light signaling is not restricted to germination and hypocotyl growth inhibition.

Together these results show that *cry1-L407F* causes strong hypersensitivity to both blue and red light in a dominant manner.

The hypersensitivity of cry1-L407F is not caused by elevated cry1 levels

Overexpression of *CRY1* under the control of the constitutive 35S promoter results in hypersensitivity of the transgenic plants towards blue light (Lin et al., 1996). Thus, we reasoned that the observed hypersensitivity of the cry1-L407F mutant could be caused by increased cry1 protein levels, although flowering time was not affected in overexpressors of wild type cry1 (Fig. 2). Quantitative immunoblots revealed, however, unchanged cry1 levels in *cry1-L407F* mutants (Fig. 7 and data not shown). These results demonstrate that the blue and red light hypersensitivity and early flowering of *cry1-L407F* is not caused by increased expression of cry1 but most likely by increased activity of the cry1-L407F protein. This conclusion is supported by the observation that blue fluence rates of up to 50 μmol m^{-2} s^{-1} caused a stronger shift of the cry1 band in the *cry1-L497F* mutant than in wild type (Fig. 7). This blue light-induced shift reflects phosphorylation of the cry1 protein that is associated with photoreceptor activation (Shalitin et al., 2002, Bouly et al., 2003, Shalitin et al., 2003). We thus conclude that the L407F mutation in cry1 increases the fraction of active (phosphorylated) photoreceptor over a broad range of blue fluence rates but has no effect in darkness as seen from the absence of a shifted cry1 band in both, wild type and the *cry1-L407F* mutant (Fig. 7).

The hypersensitivity of cry1-L407F could be caused by reduced structural flexibility of the photoreceptor

To understand potential consequences of the L407F mutation on cry1 structure and function, we carried out three independent molecular dynamics simulations for each of the four following systems: wild-type protein and the L407F mutant, both with ATP and without ATP. The time series of RMSD are useful to visualize the spatial deviation of the structure during the simulation with respect to the energy-minimized

X-ray conformation (Supplemental Fig. S3). RMSD plots of the Cα atoms show that wild-type and mutant protein are stable in all the simulated systems (RMSD = 2 – 3 Å). Generally, simulated RMSD were smaller for proteins with ATP than for proteins without ATP (Supplemental Fig. S3A). Interestingly, the crystallographic ATP binding mode was unstable in the simulations of both wild-type and mutant proteins. During the simulations of the wild-type and mutated proteins, the distance between the N6 atom of ATP and the Cγ of D409 was higher than the crystallographic distance of 3.5 Å (Supplemental Fig. S3B, top), indicating that this interaction is not maintained. Similarly, the RMSD plots of the adenine moiety of ATP showed instability of the ATP binding mode in the wild-type and mutant protein (Supplemental Fig. S3B, bottom, RMSD > 5 Å). To identify potential regions of altered structural flexibility, we plotted RMSF values of Cα atoms, which illustrate structural plasticity along the protein sequence. The RMSF plots of the wild-type protein show that ATP reduces the atomic fluctuations of three segments (Fig. 8A, top). The segments 1, 2 and 3 are spatially close to FAD and the adenine moiety of ATP, respectively (Fig. 1, left), but distant in sequence. A qualitative representation of the backbone flexibility where the thickness of the backbone is proportional to the RMSF is shown in Fig. 8B. The RMSF plots of the mutant protein show that ATP reduces the atomic fluctuations only of segment 1, because the L407F mutation has already a stabilizing influence on segments 2 and 3 (Fig. 8A).. To assess the effect of the mutation on the protein backbone flexibility, the residue-wise RMSF difference between the wild-type and mutant simulations without of ATP was calculated (Fig. 8A, bottom). The 20 residues affected by the highest flexibility reduction include the mutation site; they are located in segments 2 and 3, and are close to the three conserved tryptophans (Fig. 1C, right).

Together the simulations indicate that ATP binding stabilizes three regions of wild-type cry1 and that the L407F mutation partially mimics the effect of ATP binding by stabilizing two of three ATP binding-responsive regions even in the absence of ATP. Therefore, the differences in flexibility suggest that the L407F mutation reduces the conformational entropy penalty of ATP binding and thus might promote ATP binding, autophosphorylation and eventually cry1 signaling.

DISCUSSION

Cry1 controls flowering time

Blue light promotes flowering (Guo et al., 1998), and this effect was attributed mainly to cry2, as *cry2* mutations delay flowering under long day conditions in a phyB dependent manner (Koornneef et al., 1991, Guo et al., 1998, Guo et al., 1999, El-Assal et al., 2001, El-Din El-Assal et al., 2003, Endo et al., 2007, Liu et al., 2008). Some studies had reported that cry1 functions in flowering time regulation as well, but others failed to find such evidence (Bagnall et al., 1996, Zagotta et al., 1996, Mockler et al., 1999, Blazquez et al., 2003). While *cry2* mainly affects flowering time under long day conditions, the negative effects reported for certain *cry1* alleles and the carboxy-terminal domain of cry1 were prominent under short day conditions (Bagnall et al., 1996, Yang et al., 2001). Since both, cry1 and cry2 are involved in blue-light mediated repression of hypocotyl elongation (Ahmad et al., 1993, Lin et al., 1998), it is also possible that both act to some extent in flowering time control, but that the effect of cry1 on flowering time control is often masked by other floral regulators. In fact it was reported that *cry1 cry2* double mutants flower significantly later than *cry2* single mutants when grown under monochromatic blue light (Mockler

et al., 2003). Here, we describe the new *CRY1* allele *cry1-L407F*, which supports previous findings showing that cry1 can act as a positive regulator of the floral transition.

The *cry1-L407F* allele was isolated in a mutant screen for suppression of the late flowering phenotype of *msi1-tap1* plants. Because *cry1-L407F* and *msi1-tap1* showed an additive genetic interaction (Supplemental Fig. S1), *MSI1* and *CRY1* possibly function in independent genetic pathways of flower induction.

cry1-L407F causes increased expression of *CO* and *FT*

The cry1-L407F photoreceptor caused increased expression of *FT* (Fig. 3), and this can explain the observed early flowering phenotype. *FT* plays only a minor role in flowering time regulation under SD, but high levels of *FT* expression cause early flowering even in SD. *FT* expression is affected by light in several ways: First, light controls the phase of the circadian clock to establish the correct diurnal expression of *CO*. Second, light stabilizes CO and allows the accumulation of CO specifically under LD, when expression peak and light coincide. This coincidence of the diurnal *CO* expression peak and external light stimulus is what mainly causes photoperiodic acceleration of flowering by LD in Arabidopsis (Kobayashi et al., 2007). More recently, it was found that cry2 can also directly induce *FT* expression (Liu et al., 2008). Because increased *FT* expression in *cry1-L407F* remained constrained to the light period, it is unlikely that stabilization of CO by cry1-L407F in the dark caused the increased *FT* expression. Instead, the increased *FT* expression is probably caused by expression of *CO* during the light period in *cry1-L407F*. It is not clear how cry1-L407F caused increased CO expression during the light period. An indirect effect via altered function of the circadian clock seems unlikely given the unchanged expression

of *ELF4* and *GI* (Fig. 3). It is however conceivable that cry1-L407F directly affects CO expression, a possibility we are currently testing.

A conserved leucine is important for cry1 function

Cryptochromes are flavoproteins with two chromophores and high sequence similarity to photolyases. However, cryptochromes lack several of the characteristics of the DNA-repairing photolyases, most prominently binding to DNA, which is explained by a negative electrostatic potential of the surface around the flavin-binding pocket of DNA-photolyase (Brautigam et al., 2004, Mees et al., 2004). In addition to the amino-terminal photolyase-like PHR domain, cryptochromes contain a characteristic carboxy-terminal domain, termed CCT, which is not present in the photolyases. Expression of the CCT domain in transgenic Arabidopsis led to constitutive photomorphogenesis and mimicked the phenotype of mutations in *CONSTITUTIVE PHOTOMORPHOGENIC (COP) 1* (Yang et al., 2000). COP1 is involved in regulation of hypocotyl elongation, anthocyanin production and chloroplast development and binds to cry1 and cry2 via their CCT domains independent of light (Yang et al., 2000, Wang et al., 2001). It is possible that normally the CCT domain is kept inactive by an interaction with the PHR domain. Absorption of light would then cause conformational changes of the PHR domain leading to release of the CCT domain, which could eventually activate the signaling chain (reviewed by (Lin et al., 2005).

Here, we reported the new *cry1-L407F* allele with increased light sensitivity. In addition, the *cry1-L407F* mutants have some defects in skotomorphogenesis (cotyledons partially unfold and the light-induced *ELIP2* gene has a slightly increased basal expression in extended darkness), but there is no constitutive activation of photomorphogenesis in the dark, such as observed in plants overexpressing the CCT

domain. For example, there is no induction of cell division in the shoot apical meristem in the dark (data not shown).

The L407F amino acid substitution is within a region of the protein that is conserved in all plant crytochromes, but not in photolyases or animal cryptochromes (Fig. 1). Despite the strong conservation of leucine 407, it is not immediately obvious why the change to phenylalanine, which is of similar size and hydrophobicity like leucine, would increase the light sensitivity of cry1. The mutated L407 is close to the phosphate residues of AMP-PNP sticking out of the flavin-binding pocket in the cocrystal structure of cry1 with AMP-PNP (Brautigam et al., 2004). To test whether the L407F mutation could modulate the ATP binding to cry1, molecular dynamics simulations of cry1 wild-type and the L407F mutant in complex with FAD and with or without ATP were run. ATP binding reduced the Cα flexibility of three sequence segments, which are distant in sequence (segment 1 – 2, >50 amino acids; segment 2 – 3, >100 amino acids). In contrast to wild type, ATP binding did not reduce the flexibility of the segments 2 and 3 in the L407F mutant because their flexibility is already diminished by the single point mutation.

To explain the hyperactivity of the cry1-L407F mutant, three non mutually exclusive hypotheses can be formulated. First, the reduced flexibility of cry1-L407F could favor binding to a signaling partner, because of a reduced conformational entropy penalty upon binding. Such an effect was recently observed for a mutant of a PDZ domain (Petit et al., 2009). Second, the novel phenylalanine of the mutant is close to three conserved tryptophans, which are involved in electron transport from the surface to the FAD at least *in vitro* as extensively studied in *Escherichia coli* photolyase (Park et al., 1995, Giovani et al., 2003, Banerjee et al., 2007). Thus, it is possible that cry1-L407F has altered photochemical properties. Third, partial pre-stabilization of the ATP binding pocket in cry1-L407F could stabilize ATP-binding

extending the life time of the signaling state of cry1. This conclusion is at least consistent with the increased level of shifted and phosphorylated cry1-L407F compared to wild-type cry1 (Fig. 7). Discrimination between these possibilities will be addressed in future studies.

Hypersensitivity of cry1-L407F to various light qualities reveals tight integration of several light signaling pathways

The *cry1-L407F* allele was not only hypersensitive to blue but also to red and far-red light. Normally, red and far red light is not sensed by cryptochromes but by phytochromes. This raises the question which photoreceptor then is responsible for the increased sensitivity to red light. Because the hypersensitivity of *cry1-L407F* to red pulses could be fully reverted by far-red pulses, at least in this case it was clearly phytochrome signaling which was affected by the *cry1-L407F* allele. Interactions of red-light absorbing phytochrome and blue-light absorbing cryptochrome signaling cascades have been reported (Casal et al., 1995, Ahmad et al., 1997, Hennig et al., 1999). Furthermore, nuclear import of phyB was initiated by blue light, but not by light of 695 nm, which establishes a similar phytochrome photoequilibrium as blue light (Gil et al., 2000). Finally, cryptochromes were found to be required for phytochrome signaling to the circadian clock (Devlin et al., 2000). On the molecular level, these effects could potentially be based on an interaction of the cry1 carboxy terminal domain with phyA (Ahmad et al., 1998) or of cry2 with phyB (Mas et al., 2000). Further work will establish whether the L407F mutation in cry1 affects direct interactions with phytochrome or whether the photochemical properties of cry1-L407F are changed. We propose that the increased sensitivity of *cry1-L407F* plants to blue, red and far-red light reveals the intimate cross-talk between cryptochrome and

phytochrome light signaling cascades, which has been suggested to be important for concerted plant development under natural light conditions (Casal 2000).

MATERIALS AND METHODS

Plant material and growth conditions

Seeds of Columbia (Col) and Landsberg *erecta* (L*er*) *Arabidopsis thaliana* wild-type accessions were obtained from the Nottingham Arabidopsis Seed Stock Centre. The line *msi1-tap1* (accession Col) has been described before (Bouveret et al., 2006). The EMS allele *cry1-L407F* (accession Col) was isolated from a mutant screen (this study). To construct plants that ectopically overexpress cry1 or cry1-L407F (*35S::CRY1* and *35S::CRY1-L407F*), the full-length coding sequences were inserted into the binary destination vector pK7WG2 (Karimi et al., 2002) downstream of the cauliflower mosaic virus (CaMV) 35S promoter. Constructs were transformed into Columbia wild-type plants.

Seeds were usually germinated on sterile basal salts Murashige and Skoog (MS) medium (Duchefa, Brussels, Belgium) after two or three days stratification treatment of the imbibed seeds at 4 ºC, and plants were analyzed on plates or transferred to soil ("Einheitserde", H. Gilgen optima-Werke, Arlesheim, Switzerland) 10 days after germination. Alternatively, seeds were directly sown on soil. Plants were kept in Conviron growth chambers with mixed cold fluorescent and incandescent light (110 to 140 µmol m^{-2} s^{-1}, 21±2 ºC) under long day (LD, 16 h light) or short day (SD, 8 h light) photoperiods or were alternatively raised in green houses (LD: 14 h light, 19 ºC/10 h dark, 14 ºC; SD: 8 h light, 20 ºC/16 h dark, 20 ºC; if necessary, daylight was supplemented with mercury vapor lamps [Sylvania Lighting S.A., Meyrin, Switzerland] to a maximum of 150 µmol m^{-2} s^{-1}).

For immunoblot analyses, seeds were plated on ½ MS plates and stratified at 4°C for 4 days in darkness. Germination was induced by white light illumination for 4 h. Plants were grown at 22 °C for 4 days, and seedling were harvested after treatment with blue light emitted from LEDs (λ_{max} 471 nm, CLF Plant Climatics, Emersacker, Germany) for 30 min or 120 min with fluence rates indicated and measured with a P-2000 optometer (Gigahertz-Optik, Puchheim, Germany).

Flowering time analysis

Flowering time was scored as the length of time between the end of stratification and the development of a primary shoot of 5 mm height (= bolting). The number of rosette leaves was determined at bolting. For phase transition, all formed rosette leaves were inspected for the presence of abaxial trichomes at bolting.

RNA isolation, RT-PCR and quantitative PCR

RNA was extracted from plant tissue as previously described (Hennig et al., 2003). For RT-PCR analysis, 1 µg total RNA was treated with DNase I (Promega, Dübendorf, Switzerland). The DNA-free RNA was reverse-transcribed using a RevertAid First Strand cDNA Synthesis Kit (Fermentas, Nunningen, Switzerland) according to manufacturer's instructions. For qPCR analysis, the Universal ProbeLibrary system (Roche Diagnostics, Rotkreuz, Switzerland) was used on a 7500 Fast Real-Time PCR instrument (Applied Biosystems, Lincoln, CA). Quantitative PCR was performed with three replicates, and results were analyzed as described (Exner et al., 2009). Details of the assays used are given in supplemental Tab. S1.

Analysis of hypocotyl length

Seeds were plated on two layers of water-soaked 3 MM chromatography paper (Whatman Schleicher & Schuell, GmbH, Dassel, Germany), which were placed into clear plastic boxes. A 48 h to 96 h dark treatment at 4 ºC was followed by induction of germination by white light for 10 h (24 h for far-red studies) at 23 ºC and further incubation of the seedlings under specific light conditions, which were as follows: blue light: Philips TLD 18W/18 Blue E003, continuous light, 21 ºC; red light: Philips TLD 18W/18 Red, continuous light, 21 ºC; far red light: as described (Sperling et al., 1997), continuous light, 26 ºC. The hypocotyl length was measured by spreading the seedlings on millimetre paper and reading the length.

Quantitative immunoblots

Per sample approximately 50 seedlings were collected, frozen in liquid nitrogen and ground to a fine powder with a cell mill (MM200, Retsch, Haan, Germany). Protein was extracted by trichloracetic acid-acetone precipitation according to (Shultz et al., 2005) with the following modifications: After the washing steps, samples were dried in a SpeedVac and then dissolved in SDS sample buffer (45 mM Tris-HCl, pH 6.8; 10% glycerol; 1% SDS; 0.01% bromphenol blue; 50 mM DTT). Samples were incubated at 95 °C for 10 min followed by a centrifugation step (10 min, 20,000×g) to remove cell debris. For SDS-PAGE, 15 µg of total protein were loaded per lane on 10% SDS-minigels (Shultz et al., 2005). PageRuler™ (Fermentas, St. Leon-Rot, Germany) was used as marker. Separated proteins were transferred to nitrocellulose membranes (porablot NCP, Macherey-Nagel, Düren, Germany). Membranes were blocked with 7% milk powder in TBS (20 mM Tris-HCl, pH 7.5; 150 mM NaCl). Incubation with the two primary antibodies was done stepwise with monoclonal antibody against α-tubulin (anti-α-tubulin, produced in mouse; Clone B-5-1-2, Sigma, Taufkirchen, Germany), then with anti-cry1 antibody (raised in rabbits and provided

by M. Ahmad, Universite Paris VI, France). Both antibodies were used in a 1:2000 dilution in TBS-T (TBS with 0.1% (v/v) Tween-20). Fluorescence-labelled secondary antibodies against rabbit (donkey anti-rabbit IRDye®800CW, LI-COR Biosciences, Lincoln, Nebraska, USA) and mouse (donkey anti-mouse IRDye 700DX, Rockland, Gilbertsville, PA, USA) were incubated simultaneously for 1 h (each diluted 1:10000 in TBS-T). The membranes were scanned and analyzed with the LI-COR Odyssey Infrared Imaging System (LI-COR Biosciences, Lincoln, Nebraska, USA). Bands detected in the 700 nm channel correspond to α-tubulin, bands detected in the 800 nm channel to cry1. The system was calibrated to ensure measurements in the linear range for both, α-tubulin and cry1. The cry1 signal was normalized against the α-tubulin signal. In addition, the percentage of shifted cry1 bands compared to the total cry1 signal was determined.

Sequence alignment

The Arabidopsis CRY1 (At4g08920), CRY2 (At1g04400) and photolyase (At3g15620) protein sequences were obtained from TAIR (http://www.arabidopsis.org/) and blasted against the non-redundant protein data bases at NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=Search&db=pubmed) and at the DOE Joint Genome Institute (http://genome.jgi-psf.org/). The obtained sequences were aligned using ClustalX 2.0. The identifiers of the protein sequences included in this analysis are listed in supplemental Table S2. Nomenclature of phytochrome as well as cryptochrome photoreceptors and their genes is according to (Quail et al., 1994).

Molecular Dynamics Simulations

The crystal structure of the PHR domain of Arabidopsis cry1 with AMPPNP bound (PDB accession code: 1U3D) was used for modeling and molecular dynamics simulations. The L407F mutation was introduced with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r1, Schrödinger, LLC) and the most common rotamer was selected. To generate ATP from AMP-PNP, the nitrogen atom between phosphate groups of AMP-PNP was replaced by oxygen. Ions and crystallization water were kept for further calculations. All the simulations were carried out using CHARMM (Brooks et al., 1983) (version c35b2) and the PARAM22 force field (MacKerell et al., 1998, Mackerell, Jr. et al., 2004) with the TIP3P model of water (Jorgensen et al., 1983, MacKerell et al., 1998). To effectively compare simulations with experiments, pH of 7.4 was considered. The side chains of aspartates and glutamates were negatively charged, those of lysines and arginines were positively charged, histidines were considered neutral, the N-terminus was positively charged and the C-terminus negatively charged. First, structures were minimized *in vacuo* using a dielectric constant $\varepsilon=4r$ (where r is the distance in Å between atoms/partial charges) to an energy gradient of 0.01 Kcal mol^{-1} $Å^{-1}$. The minimized protein was then inserted into a water box where each atom of the protein had a distance of at least 14 Å from the boundary. Water molecules within 2.8 Å from any atom of the protein were removed. Chloride and sodium ions were added to neutralize the total charge of the system at a concentration of 200 mM. The final system consisted of around 96000 atoms, circa 7900 of which belong to the solute. To avoid finite-size effects, periodic boundary conditions were applied. Long-range electrostatic effects were taken into account by the Particle Mesh Ewald summation method (Darden et al., 1993). The temperature was kept constant at 300 K by the Nosé-Hoover thermostat (Nose 1984, Hoover 1985), while the pressure was held constant at 1 atm by applying the Langevin piston pressostat. Lookup tables (Nilsson 2009) for the calculation of water-

water nonbonded interactions (van der Waals and Coulomb) were used to increase efficiency. SHAKE was applied to the hydrogens allowing an integration step of 2 fs. Different initial random velocities were assigned to every simulation. Four systems were simulated: the wild-type protein with only FAD bound; the wild-type protein with FAD and ATP bound; the mutant with only FAD bound; and the mutant with FAD and ATP bound. Three independent 30-ns long simulations were carried out for each system.

Trajectory Analyses

Root mean square deviations and root mean square fluctuations, RMSD and RMSF respectively, were calculated with CHARMM and their formula is:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - x_{i,ref}\right)^2 + \left(y_i - y_{i,ref}\right)^2 + \left(z_i - z_{i,ref}\right)^2}$$

$$RMSF_i = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N_f}\left(x_i - x_{i,ave}\right)^2 + \left(y_i - y_{i,ave}\right)^2 + \left(z_i - z_{i,ave}\right)^2}$$

where N is the number of atoms; x_i, y_i, z_i are the coordinates of the atom i after best superposition on a reference structure; $x_{i,ref}$, $y_{i,ref}$, $z_{i,ref}$ are the coordinates of the atom i in the reference structure; the coordinates x_i, y_i, z_i refer to the average structure; N_f is the number of frames in the trajectory segment analyzed for RMSF calculations; the coordinates $x_{i,ave}$, $y_{i,ave}$, $z_{i,ave}$ refer to the average structure. The reference structure for RMSD analyses was the starting structure used in the dynamics, i.e. the energy-minimized X-ray structure. The average structures and RMSF were calculated along 2 ns segments of trajectory, skipping the first 2 ns and the last incomplete segment shorter than 2 ns. For the first 30 ns of simulation time, 13 values of RMSF were therefore calculated and then averaged. RMSD expresses how different an object is with respect to another after the best superposition of the two. A RMSD value of zero

means perfect superposition. RMSF is a measure of atomic flexibility and it can be related to the crystallographic B-factor, $B = 8\pi^2 / 3(RMSF)^2$. The distance between the N6 atom of ATP and the C_γ of D409 was calculated with the program Wordom (Seeber et al., 2007). Structures were plotted with PyMol.

ACKNOWLEDGEMENTS

We thank Romaric Bouveret for help during the EMS treatment of *msi1-tap1* seeds. This work was supported by SNF grant 3100AO-116060 (to L.H.), ETH project TH-16/05-2 (to L.H.), and DFG grant BA985/11-1 (to A.B.).

REFERENCES

Ahmad M, Cashmore AR (1993) *Hy4* gene of *Arabidopsis thaliana* encodes a protein with characteristics of a blue-light photoreceptor. Nature **366**: 162-166

Ahmad M, Cashmore AR (1997) The blue-light receptor cryptochrome 1 shows functional dependence on phytochrome A or phytochrome B in *Arabidopsis thaliana*. Plant J **11**: 421-427

Ahmad M, Grancher N, Heil M, Black RC, Giovani B, Galland P, Lardemer D (2002) Action spectrum for cryptochrome-dependent hypocotyl growth inhibition in Arabidopsis. Plant Physiol **129**: 774-785

Ahmad M, Jarillo JA, Smirnova O, Cashmore AR (1998) The cry1 blue light photoreceptor of Arabidopsis interacts with phytochrome A in vitro. Mol Cell **1**: 939-948

Bagnall DJ, King RW, Hangarter RP (1996) Blue-light promotion of flowering is absent in *hy4* mutants of Arabidopsis. Planta **200**: 278-280

Banerjee R, Batschauer A (2005) Plant blue-light receptors. Planta **220**: 498-502

Banerjee R, Schleicher E, Meier S, Viana RM, Pokorny R, Ahmad M, Bittl R, Batschauer A (2007) The signaling state of Arabidopsis cryptochrome 2 contains flavin semiquinone. J Biol Chem **282**: 14916-14922

Blazquez MA, Ahn JH, Weigel D (2003) A thermosensory pathway controlling flowering time in *Arabidopsis thaliana*. Nat Genet **33**: 168-171

Bouly JP, Giovani B, Djamei A, Mueller M, Zeugner A, Dudkin EA, Batschauer A, Ahmad M (2003) Novel ATP-binding and autophosphorylation activity associated with Arabidopsis and human cryptochrome-1. Eur J Biochem **270**: 2921-2928

Bouveret R, Schönrock N, Gruissem W, Hennig L (2006) Regulation of flowering time by Arabidopsis MSI1. Development **133**: 1693-1702

Brautigam CA, Smith BS, Ma Z, Palnitkar M, Tomchick DR, Machius M, Deisenhofer J (2004) Structure of the photolyase-like domain of cryptochrome 1 from *Arabidopsis thaliana*. Proc Natl Acad Sci U S A **101**: 12142-12147

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem **4**: 187-217

Casal JJ (2000) Phytochromes, cryptochromes, phototropin: photoreceptor interactions in plants. Photochem Photobiol **71**: 1-11

Casal JJ, Boccalandro H (1995) Co-action between phytochrome B and HY4 in *Arabidopsis thaliana*. Planta **197**: 213-218

Casal JJ, Mazzella MA (1998) Conditional synergism between cryptochrome 1 and phytochrome B is shown by the analysis of *phya*, *phyb*, and *hy4* simple, double, and triple mutants in Arabidopsis. Plant Physiol **118**: 19-25

Darden T, York D, Pedersen L (1993) An N· log (N) method for Ewald sums in large systems. J Chem Phys **98**: 10089-10092

Devlin PF, Kay SA (2000) Cryptochromes are required for phytochrome signaling to the circadian clock but not for rhythmicity. Plant Cell **12**: 2499-2510

Doyle MR, Davis SJ, Bastow RM, H. G, McWatters, Kozma-Bognar L, Nagy F, Millar AJ, Amasino RM (2002) The *ELF4* gene controls circadian rhythms and flowering time in *Arabidopsis thaliana*. Nature **419**: 74-77

El-Assal SE, Alonso-Blanco C, Peeters AJ, Raz V, Koornneef M (2001) A QTL for flowering time in Arabidopsis reveals a novel allele of cry2. Nat Genet **29**: 435-440

El-Din El-Assal S, Alonso-Blanco C, Peeters AJ, Wagemaker C, Weller JL, Koornneef M (2003) The role of cryptochrome 2 in flowering in Arabidopsis. Plant Physiol **133**: 1504-1516

Endo M, Mochizuki N, Suzuki T, Nagatani A (2007) CRYPTOCHROME2 in vascular bundles regulates flowering in Arabidopsis. Plant Cell **19**: 84-93

Exner V, Aichinger E, Shu H, Wildhaber T, Alfarano P, Caflisch A, Gruissem W, Köhler C, Hennig L (2009) The chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 is essential for H3K27me3 binding and function during Arabidopsis development. PLoS ONE **4**: e5335

Franklin KA, Larner VS, Whitelam GC (2005) The signal transducing photoreceptors of plants. Int J Dev Biol **49**: 653-664

Gil P, Kircher S, Adam E, Bury E, Kozma-Bognar L, Schäfer E, Nagy F (2000) Photocontrol of subcellular partitioning of phytochrome-B:GFP fusion protein in tobacco seedlings. Plant J **22**: 135-145

Giovani B, Byrdin M, Ahmad M, Brettel K (2003) Light-induced electron transfer in a cryptochrome blue-light photoreceptor. Nat Struct Biol **10**: 489-490

Guo H, Duong H, Ma N, Lin C (1999) The Arabidopsis blue light receptor cryptochrome 2 is a nuclear protein regulated by a blue light-dependent post-transcriptional mechanism. Plant J **19**: 279-287

Guo HW, Yang WY, Mockler TC, Lin CT (1998) Regulations of flowering time by Arabidopsis photoreceptors. Science **279**: 1360-1363

Harari-Steinberg O, Ohad I, Chamovitz DA (2001) Dissection of the light signal transduction pathways regulating the two early light-induced protein genes in Arabidopsis. Plant Physiol **127**: 986-997

Hennig L, Funk M, Whitelam GC, Schäfer E (1999) Functional interaction of cryptochrome 1 and phytochrome D. Plant J. **20**: 289-294

Hennig L, Taranto P, Walser M, Schönrock N, Gruissem W (2003) Arabidopsis MSI1 is required for epigenetic maintenance of reproductive development. Development **130**: 2555-2565

Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. Phys Rev A **31**: 1695-1697

Johnson E, Bradley M, Harberd NP, Whitelam GC (1994) Photoresponses of light-grown *phyA* mutants of Arabidopsis. Plant Physiol **105**: 141-149

Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein MK (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys **79**: 926-935

Josse EM, Foreman J, Halliday KJ (2008) Paths through the phytochrome network. Plant Cell Environ **31**: 667-678

Karimi M, Inze D, Depicker A (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci **7**: 193-195

Kobayashi Y, Weigel D (2007) Move on up, it's time for change--mobile signals controlling photoperiod-dependent flowering. Genes Dev **21**: 2371-2384

Koornneef M, Hanhart CJ, J. H. van der Veen (1991) A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol Gen Genet **229**: 57-66

Koornneef M, Rolff E, Spruit CJP (1980) Genetic control of light-inhibited hypocotyl elongation in *Arabidopsis thaliana*. Z Pflanzenphysiol **100**: 147-160

Lin C, Ahmad M, Gordon D, Cashmore A (1995) Expression of an Arabidopsis cryptochrome gene in transgenic tobacco results in hypersensitivity to blue, UV-A, and green light. Proc Natl Acad Sci USA **92**: 8423-8427

Lin C, Todo T (2005) The cryptochromes. Genome Biol **6**: 220

Lin CT, Ahmad M, Cashmore AR (1996) Arabidopsis cryptochrome 1 is a soluble protein mediating blue light-dependent regulation of plant growth and development. Plant Journal **10**: 893-902

Lin CT, Yang HY, Guo HW, Mockler T, Chen J, Cashmore AR (1998) Enhancement of blue-light sensitivity of Arabidopsis seedlings by a blue light receptor cryptochrome 2. Proc Natl Acad Sci USA **95**: 2686-2690

Liu H, Yu X, Li K, Klejnot J, Yang H, Lisiero D, Lin C (2008) Photoexcited CRY2 interacts with CIB1 to regulate transcription and floral initiation in Arabidopsis. Science **322**: 1535-1539

MacKerell AD, Bashford D, Bellot M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen T, Prodhom B (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B **102**: 3586-3616

Mackerell AD, Jr., Feig M, Brooks CL 3r (2004) Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem **25**: 1400-1415

Mas P, Devlin PF, Panda S, Kay SA (2000) Functional interaction of phytochrome B and cryptochrome 2. Nature **408**: 207-211

Mees A, Klar T, Gnau P, Hennecke U, Eker AP, Carell T, Essen LO (2004) Crystal structure of a photolyase bound to a CPD-like DNA lesion after in situ repair. Science **306**: 1789-1793

Mockler T, Yang H, Yu X, Parikh D, Cheng YC, Dolan S, Lin C (2003) Regulation of photoperiodic flowering by Arabidopsis photoreceptors. Proc Natl Acad Sci U S A **100**: 2140-2145

Mockler TC, Guo HW, Yang HY, Duong H, Lin CT (1999) Antagonistic actions of Arabidopsis cryptochromes and phytochrome B in the regulation of floral induction. Development **126**: 2073-2082

Muller M, Carell T (2009) Structural biology of DNA photolyases and cryptochromes. Curr Opin Struct Biol **19**: 277-285

Neff MM, Chory J (1998) Genetic interactions between phytochrome A, phytochrome B, and cryptochrome 1 during Arabidopsis development. Plant Physiol **118**: 27-36

Nilsson L (2009) Efficient table lookup without inverse square roots for calculation of pair wise atomic interactions in classical simulations. J Comput Chem **30**: 1490-1498

Nose S (1984) A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys **81**: 511-520

Oh E, Yamaguchi S, Hu J, Yusuke J, Jung B, Paik I, Lee HS, Sun TP, Kamiya Y, Choi G (2007) PIL5, a phytochrome-interacting bHLH protein, regulates gibberellin responsiveness by binding directly to the *GAI* and *RGA* promoters in Arabidopsis seeds. Plant Cell **19**: 1192-1208

Park DH, Somers DE, Kim YS, Choy YH, Lim HK, Soh MS, Kim HJ, Kay SA, Nam HG (1999) Control of circadian rhythms and photoperiodic flowering by the Arabidopsis *GIGANTEA* gene. Science **285**: 1579-1582

Park HW, Kim ST, Sancar A, Deisenhofer J (1995) Crystal structure of DNA photolyase from *Escherichia coli*. Science **268**: 1866-1872

Petit CM, Zhang J, Sapienza PJ, Fuentes EJ, Lee AL (2009) Hidden dynamic allostery in a PDZ domain. Proc Natl Acad Sci U S A **106**: 18249-18254

Pokorny R, Klar T, Hennecke U, Carell T, Batschauer A, Essen LO (2008) Recognition and repair of UV lesions in loop structures of duplex DNA by DASH-type cryptochrome. Proc Natl Acad Sci U S A **105**: 21023-21027

Quail PH, Briggs WR, Chory J, Hangarter RP, Harberd NP, Kendrick RE, Koornneef M, Parks B, Sharrock RA, Schäfer E, Thompson WF, Whitelam GC (1994) Spotlight on phytochrome nomenclature. Plant Cell **6**: 468-471

Reed JW, Nagatani A, Elich TD, Fagan M, Chory J (1994) Phytochrome A and phytochrome B have overlapping but distinct functions in Arabidopsis development. Plant Physiol **104**: 1139-1149

Reed JW, Nagpal P, Poole DS, Furuya M, Chory J (1993) Mutations in the gene for the red-far-red light receptor phytochrome B alter cell elongation and physiological responses throughout Arabidopsis development. Plant Cell **5**: 147-157

Seeber M, Cecchini M, Rao F, Settanni G, Caflisch A (2007) Wordom: a program for efficient analysis of molecular dynamics simulations. Bioinformatics **23**: 2625-2627

Selby CP, Sancar A (2006) A cryptochrome/photolyase class of enzymes with single-stranded DNA-specific photolyase activity. Proc Natl Acad Sci U S A **103**: 17696-17700

Shalitin D, Yang H, Mockler TC, Maymon M, Guo H, Whitelam GC, Lin C (2002) Regulation of Arabidopsis cryptochrome 2 by blue-light-dependent phosphorylation. Nature **417**: 763-767

Shalitin D, Yu X, Maymon M, Mockler T, Lin C (2003) Blue light-dependent in vivo and in vitro phosphorylation of Arabidopsis cryptochrome 1. Plant Cell **15**: 2421-2429

Shinomura T, Nagatani A, Hanzawa H, Kubota M, Watanabe M, Furuya M (1996) Action spectra for phytochrome A- and B-specific photoinduction of seed germination in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **93**: 8129-8133

Shultz RW, Settlage RE, Hanley-Bowdoin L, Thompson WF (2005) A trichloroacetic acid-acetone method greatly reduces infrared autofluorescence of protein extracts from plant tissue. Plant Mol Biol Rep **23**: 405-409

Sperling U, van Cleve B, Frick G, Apel K, Armstrong GA (1997) Overexpression of light-dependent PORA or PORB in plants depleted of endogenous POR by far-red light enhances seedling survival in white light and protects against photooxidative damage. Plant J **12**: 649-658

Wang H, Ma LG, Li JM, Zhao HY, Deng XW (2001) Direct interaction of Arabidopsis cryptochromes with COP1 in light control development. Science **294**: 154-158

Yang HQ, Tang RH, Cashmore AR (2001) The signaling mechanism of Arabidopsis cry1 involves direct interaction with cop1. Plant Cell **13**: 2573-2587

Yang HQ, Wu YJ, Tang RH, Liu D, Liu Y, Cashmore AR (2000) The C termini of Arabidopsis cryptochromes mediate a constitutive light response. Cell **103**: 815-827

Yu X, Shalitin D, Liu X, Maymon M, Klejnot J, Yang H, Lopez J, Zhao X, Bendehakkalu KT, Lin C (2007) Derepression of the NC80 motif is critical for the photoactivation of Arabidopsis CRY2. Proc Natl Acad Sci U S A **104**: 7289-7294

Zagotta MT, Hicks KA, Jacobs CI, Young JC, Hangarter RP, Meekswagner DR (1996) The Arabidopsis *ELF3* gene regulates vegetative photomorphogenesis and the photoperiodic induction of flowering. Plant Journal **10**: 691-702

TABLES

Table 1. Flowering time of *03 457*.

Shown are mean ± S.E. (n ≥ 10). Note that 0.3 457 was in the *msi1-tap1* background.

genotype	flowering time in LD (days)	flowering time in SD (days)
Col	23.9 ± 0.6	81.8 ± 1.8
msi1-tap1	38.5 ± 1.2	117.9 ± 6.1
0.3 457	28.4 ± 0.3	51.4 ± 1.7

FIGURE LEGENDS

Figure 1. CRY1 sequence context of the conserved leucine 407. A: Structure of the Arabidopsis *CRY1* gene. Boxes represent exons; positions of translational start and stop as well as of the L407F mutation are shown; grey boxes represent untranslated regions. **B:** Cryptochrome and photolyase protein sequences of several organisms were aligned, and a segment of the generated sequence alignment is shown; numbers at the end of each line indicate the amino acid position within the respective protein. The arrow head marks the leucine that is exchanged for a phenylalanine in CRY1-L407F. Note that this leucine is usually conserved in plant cryptochromes, but not in photolyases. The arrows indicate conserved tryptophans involved in electron transfer to FAD. The grey box marks a plant cryptochrome-specific 12 amino acid motif, which includes L407. **C:** X-ray structure of the complex of cry1, FAD and AMP-PNP taken from (Brautigam et al., 2004) (Left). Conserved Trp residues are in black, and L407 is in magenta. The red, yellow and orange sequence regions (labeled 1, 2 and 3, respectively) correspond to the peaks in the RMSF plot (see main text). Structural model of L407F cry1 (Right). The mutation L407F is in magenta. The first 20 residues, for which the reduction of flexibility caused by the L407F mutation is highest, are in yellow.

Figure 2. cry1-L407F mediates early flowering. A: Juvenile-adult phase transition of Col wild-type (WT) and *cry1-L407F* plants under long day photoperiods. **B:** Flowering time of Col wild-type and *cry1-L407F* plants under long day (LD) and short day (SD) photoperiods. **C:** Flowering time of wild-type plants and four randomly selected transgenic cry1 or cry1-L407F overexpressing lines (OE) under short day photoperiods. Diagrams show means and standard errors ($n \geq 14$).

Figure 3. *CO* and *FT* transcript levels are increased in the *cry1-L407F* mutant. Quantitative RT-PCR was performed on cDNA from 15 day-old seedlings grown under short day conditions. Relative expression values are shown as mean and standard error ($n \geq 4$). White and grey bars on top of the diagrams represent periods of light and darkness, respectively. Values were normalized to a *PP2A* gene (At1g13320).

Figure 4. Inhibition of hypocotyl elongation in *cry1-L407F* mutant seedlings is hypersensitive to blue, red and far-red light. Fluence rate response curves of hypocotyl growth inhibition under continuous light treatments. Diagrams show means and standard errors of three replicates with at least 15 seedlings analyzed in each experiment. Wild type (WT), filled circles; *cry1-L407F*, open circles.

Figure 5. Induction of germination in *cry1-L407F* mutants is hypersensitive to white and red light. Seeds were sown under green light, stratified for 2 days at 4°C and treated with light pulses. After 4 days in the dark, germination rates were determined. DD, no light treatment; pR, 30 min of red light (10.5 μmol m^{-2} s^{-1}); pW, 30 min of white light (130 μmol m^{-2} s^{-1}); pFR, 30 min of far red light (110 μmol m^{-2} s^{-1}); pR-pFR, red pulse followed by far red pulse; cW, continuous white light for 4 days (130 μmol m^{-2} s^{-1}). Shown are means and standard errors of 4 replicate experiments with three different seed batches.

Figure 6. Induction of *ELIP2* expression in *cry1-L407F* mutants is hypersensitive to red light. Seedlings were grown for 4 days in the dark before transferred to continuous red light (10.5 μmol m^{-2} s^{-1}). Relative expression values based on RT-

qPCR are shown as mean and standard error ($n \geq 3$). Values were normalized to a *PP2A* gene (At1g13320).

Figure 7. Cry1-L407 is stronger phosphorylated in blue light than wild-type cry1. Immunoblot analysis of wild-type cry1 (cry1-WT) and cry1-L407F protein levels (both in *msi1-tap1* background). Seedlings were grown for 96 hours in complete darkness before transferred to monochromatic blue light (λ_{max} 471 nm) of given fluence rates. **A:** Representative immunoblot of samples kept in darkness for 96 hours (0) and then irradiated with blue light for 30 min or 120 min, or kept in darkness for another 120 min (120d). The blot shows samples from irradiation with 25 μmol m^{-2} s^{-1} blue light. The cry1 and tubulin signals are indicated with arrows. Note the shifted cry1 bands appearing in the light-treated samples that correspond to phosphorylated forms of cry1. **B:** Ratios of shifted (phosphorylated) cry1 to total amount of cry1 in the respective genotypes. Seedlings were irradiated with the given fluence rates of blue light for 30 min. **C:** Same as B but seedlings were treated for 120 min with blue light. Quantification of the bands was done with the LIC-COR Odyssey infra-red imaging system and software in the linear range of cry1 and tubulin signals.

Figure 8. The L407F mutation reduces structural flexibility of cry1. A: Comparison of backbone flexibility. Each curve is the average of the RMSF calculated over three trajectories. The segments corresponding to the peaks labeled 1, 2, and 3 are shown in Fig. 1. **B:** The tube-like rendering of the backbone flexibility of wild-type (left) and L407F cry1 (right) was generated using values from (A).

Supplement

Supplementary Table S1. Primers used for quantitative RT PCR.

Gene	Forward primer	Reverse primer	Universal ProbeLibrary probe
FT, At1g65480	GGTGGAGAAGACCTCAGGAA	GGTTGCTAGGACTTGGAACATC	#138 (Arabidopsis)
ELF4, At2G40080	AGTTTCTCGTCGGGCTTTC	GCTCTAGTTCCGGCAGCA	#157 (Arabidopsis)
GI, At1G22770	TTTCCGATGGTGTAGTGGTG	TTGAAGGCATCAGTTGAGGA	#67 (Arabdopsis)
CO, At5g15840	GCCTACTTGTGCATGAGCTG	GTTTATGGCGGGAAGCAAC	#53 (Arabidopsis)
ELIP2, At4G14690	CCACCACAAATGCCACAG	GCAAATCTCCAAACTTCGTACTC	#101 (Arabidopsis)
PP2A, At1g13320[1)]	GGAGAGTGACTTGGTTGAGCA	CATTCACCAGCTGAAAGTCG	#82 (Arabidopsis)

[1)] Reference gene.

Supplementary Table S2. Protein sequences used for the alignment.

Organism	Label	Identifier
Arabidopsis thaliana	At_CRY1	GI: 826470
Arabidopsis thaliana	At_CRY2	GI: 839529
Arabidopsis thaliana	At_Photolyase	GI: 828632
Danio rerio	Dr_CRY1	GI:8698584
Drosophila melanogaster	Dm_CRY1	GI:3986298
Physcomitrella patens	Pp_CRY1	jgi|Phypa1_1|111603|e_gw1.2.5.1
Physcomitrella patens	Pp_Photolyase	jgi|Phypa1_1|200971|estExt_gwp_gw1.C_4880009
Populus trichocarpa	Pt_CRY1	estExt_fgenesh4_pm.C_LG_II0442
Populus trichocarpa	Pt_CRY2	gw1.273.26.1
Populus trichocarpa	Pt_Photolyase	fgenesh4_pg.C_LG_III000326
Oryza sativa	Os_CRY1	GI: 115458700
Oryza sativa	Os_CRY2	GI: 28372347
Oryza sativa	Os_Photolyase	GI:125581224

Supplementary Figure S1. *cry1-L407F* and *msi1-tap1* affect flowering time additively.

cry1-L407F msi1-tap1 flowers later than *cry1-L407F* but earlier than *msi1-tap1* under long and short day conditions. **A:** Rosette leaf number at bolting under long day conditions. **B:** Phase transition under long day experiments. Diagrams represent means and standard errors ($n \geq 10$); asterisks indicate student's t-test results: single asterisk = $p<0.05$, double asterisk = $p<0.005$, black: test against Col, grey: test against *cry1-L407F*, red: test against *msi1-tap1*. **C:** 45 days old plants grown under short day conditions; scale bar: 1 cm.

Supplementary Figure S2. Emission spectrum of the red light source.

Supplementary Figure S3.

A: Structural stability of the PHR domain of cry1. The Cα atoms RMSD of residues 13-497 was calculated using the energy-minimized X-ray structure (PDB accession code 1U3D) as reference. Each color represents an independent MD run.

B: Stability of the ATP binding mode. (Top) The dashed line shows the distance ATP/N6 – D409/Cγ in the X-ray structure (3.5 Å). (Bottom) The RMSD of the adenine heavy atoms was calculated upon optimal overlap of the Cα atoms of the PHR domain of cry1. Only in one wild type simulation (blue time series, bottom left), the position of ATP was almost conserved, but the interaction between D409 and the N6 atom of ATP was not established (blue time series in Fig. S3B, top left).

Additionally, in one mutant simulation (blue time series, top right and bottom right), the ATP assumed a different binding mode in which the adenine moiety and D409 side chains were more exposed to solvent but at the same hydrogen bond distance as in the X-ray structure (see Fig. S4) of the N6 atom of ATP and the Cγ of D409.

Supplementary Figure S4. Binding mode displacement of ATP. Energy-minimized X-ray structure (green) and last snapshot from the molecular dynamics (red). In one mutant simulation (blue time series in Fig. S3B, top right and bottom right), the ATP rebinds with a different orientation of the adenine and ribose groups after about 22 ns. The hydrogen bond between the N6 atom of ATP and the Cγ of D409 is preserved.

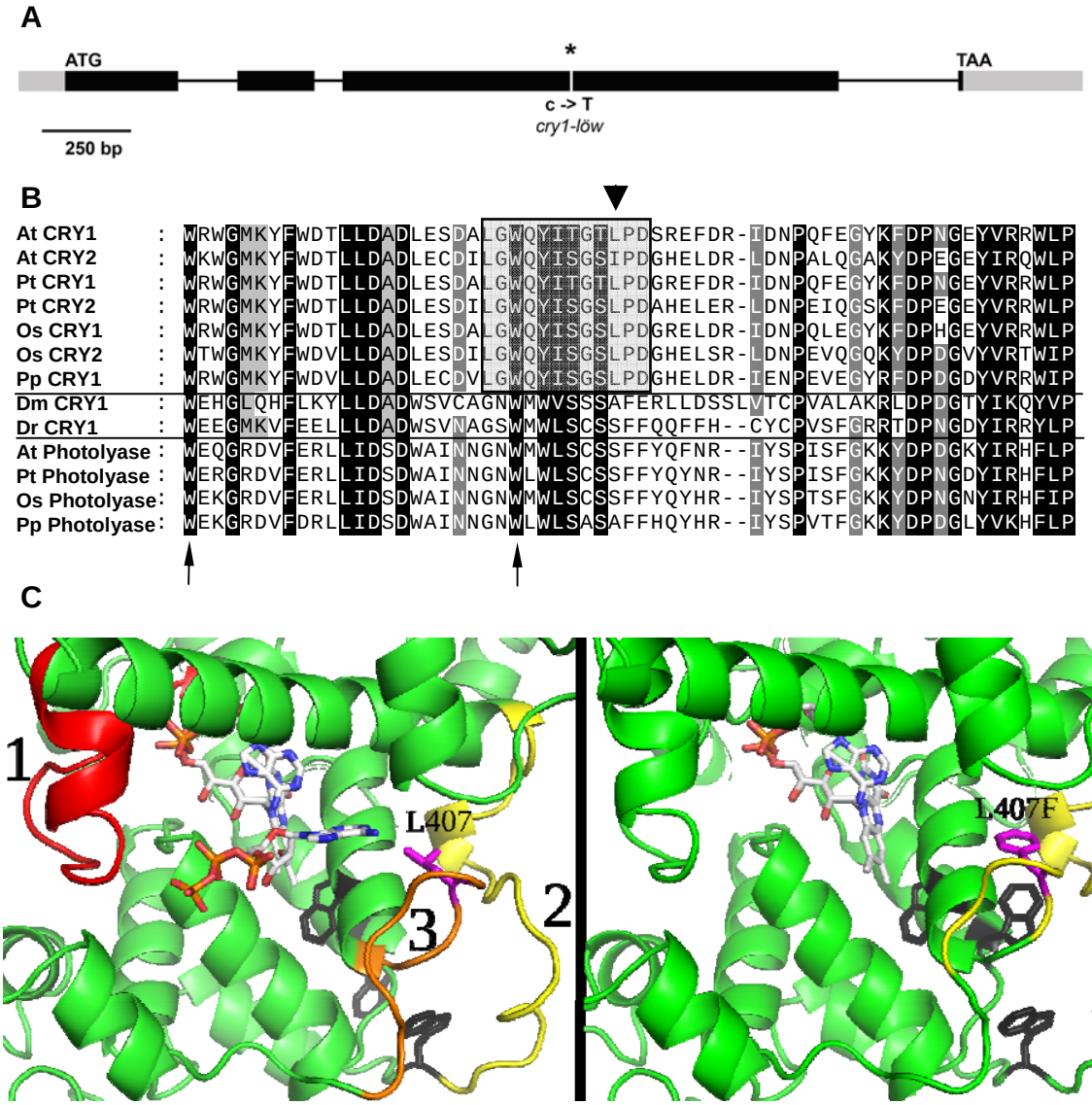

A
ATG
*
TAA
c -> T
cry1-löw
250 bp
B
At CRY1 : WRWGMKYFWDTLLDADLESDALGWQYITGTLPDSREFDR-IDNPQFEGYKFDPNGEYVRRWLP : 438
At CRY2 : WKWGMKYFWDTLLDADLECDILGWQYISGSIPDGHELDR-LDNPALQGAKYDPEGEYIRQWLP : 435
Pt CRY1 : WRWGMKYFWDTLLDADLESDALGWQYITGTLPDGREFDR-IDNPQFEGYKFDPNGEYVRRWLP : 431
Pt CRY2 : WRWGMKYFWDTLLDADLESDILGWQYISGSLPDAHELER-LDNPEIQGSKFDPEGEYVRRWLP : 430
Os CRY1 : WRWGMKYFWDTLLDADLESDALGWQYISGSLPDGRELDR-IDNPQLEGYKFDPHGEYVRRWLP : 446
Os CRY2 : WTWGMKYFWDVLLDADLESDILGWQYISGSLPDGHELSR-LDNPEVQGQKYDPDGVYVRTWIP : 435
Pp CRY1 : WRWGMKYFWDVLLDADLECDVLGWQYISGSLPDGHELDR-IENPEVEGYRFDPDGDYVRRWIP : 434
Dm CRY1 : WEHGLQHFLKYLLDADWSVCAGNWMWVSSSAFERLLDSSLVTCPVALAKRLDPDGTYIKQYVP : 459
Dr CRY1 : WEEGMKVFEELLLDADWSVNAGSWMWLSCSSFFQQFFH--CYCPVSFGRRTDPNGDYIRRYLP : 434
At Photolyase : WEQGRDVFERLLIDSDWAINNGNWMWLSCSSFFYQFNR--IYSPISFGKKYDPDGKYIRHFLP : 462
Pt Photolyase : WERGRDVFERLLIDSDWAINNGNWLWLSCSSFFYQYNR--IYSPISFGKKYDPNGDYIRHFLP : 443
Os Photolyase : WEKGRDVFERLLIDSDWAINNGNWMWLSCSSFFYQYHR--IYSPTSFGKKYDPNGNYIRHFIP : 463
Pp Photolyase : WEKGRDVFDRLLIDSDWAINNGNWLWLSASAFFHQYHR--IYSPVTFGKKYDPDGLYVKHFLP : 449
C
1
L407
3
2
L407F

Fig. 1

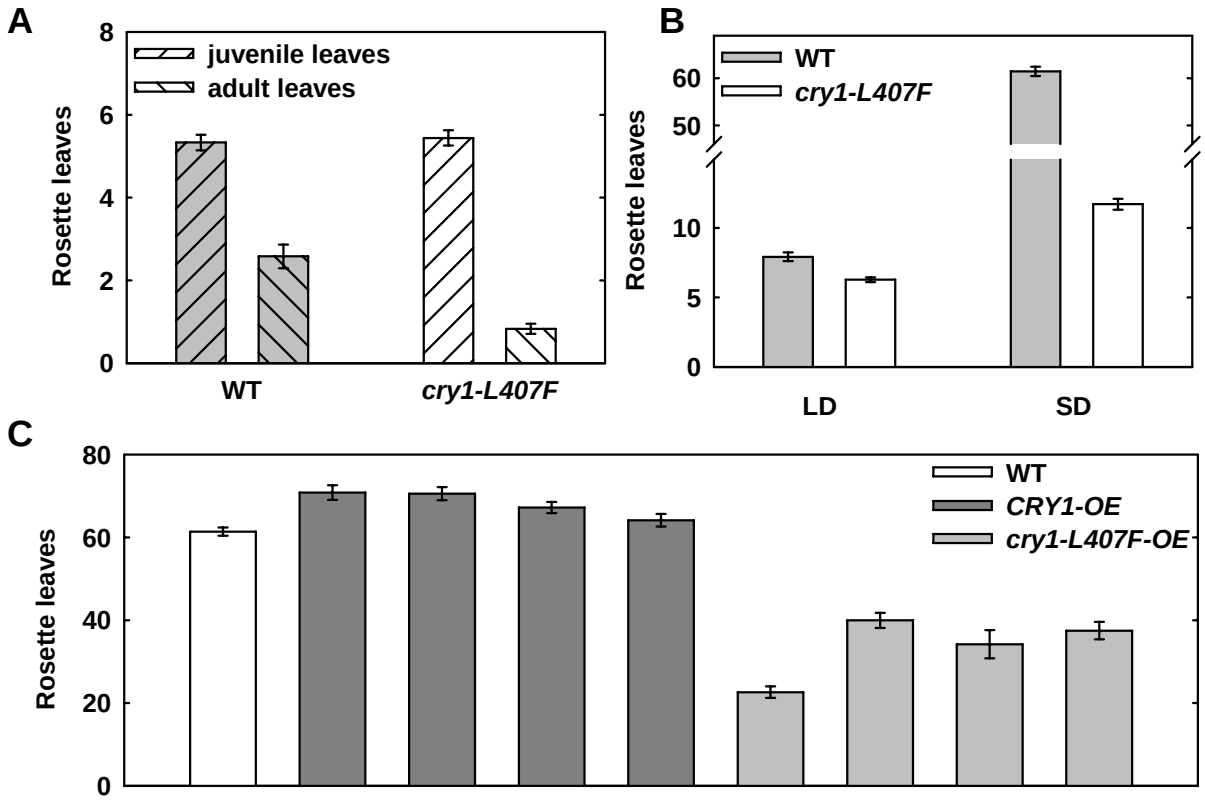

A
Rosette leaves
juvenile leaves
adult leaves
WT
cry1-L407F
B
Rosette leaves
WT
cry1-L407F
LD
SD
C
Rosette leaves
WT
CRY1-OE
cry1-L407F-OE

Fig. 2

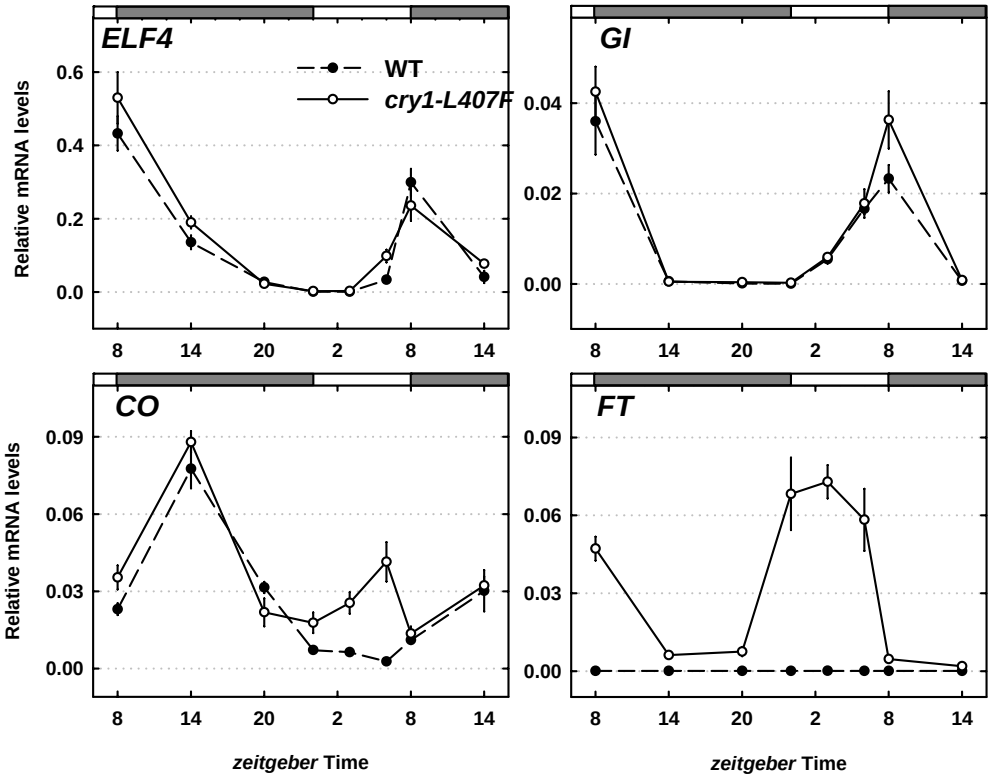

ELF4
WT
cry1-L407F
GI
CO
FT
Relative mRNA levels
Relative mRNA levels
0.6
0.4
0.2
0.0
0.04
0.02
0.00
0.09
0.06
0.03
0.00
8
14
20
2
8
14
zeitgeber Time
zeitgeber Time

Fig. 3

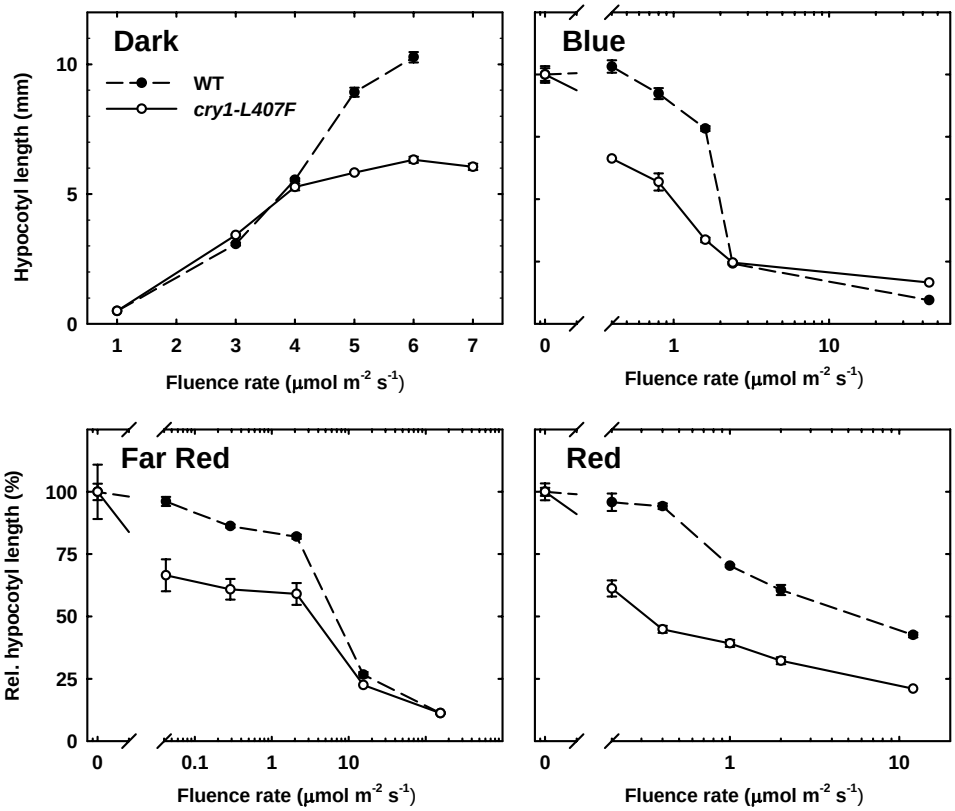

Dark
WT
cry1-L407F
Hypocotyl length (mm)
Fluence rate (µmol m-2 s-1)
Blue
Fluence rate (µmol m-2 s-1)
Far Red
Rel. hypocotyl length (%)
Fluence rate (µmol m-2 s-1)
Red
Fluence rate (µmol m-2 s-1)

Fig. 4

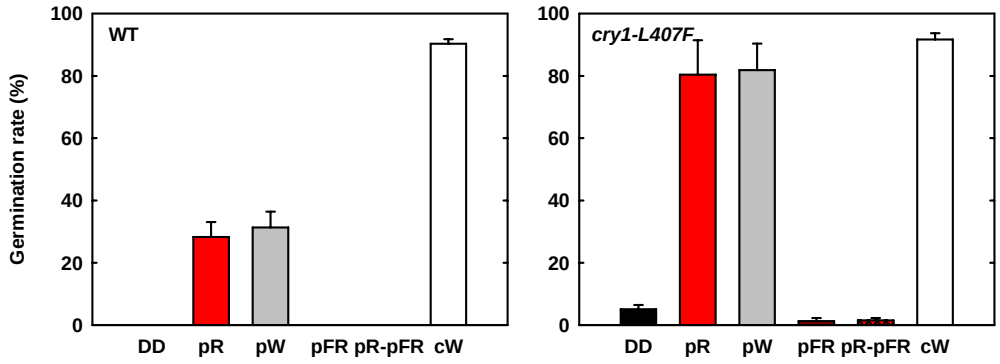

Germination rate (%)
100
80
60
40
20
0
WT
DD
pR
pW
pFR
pR-pFR
cW
cry1-L407F

Fig. 5

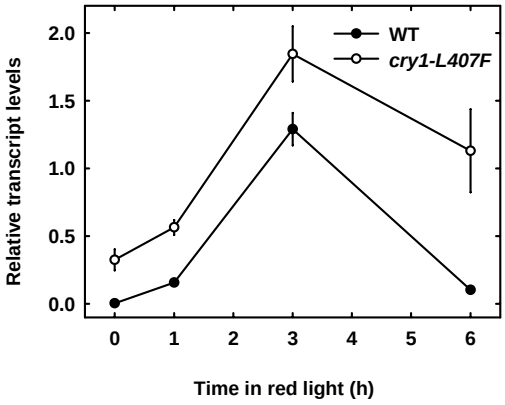
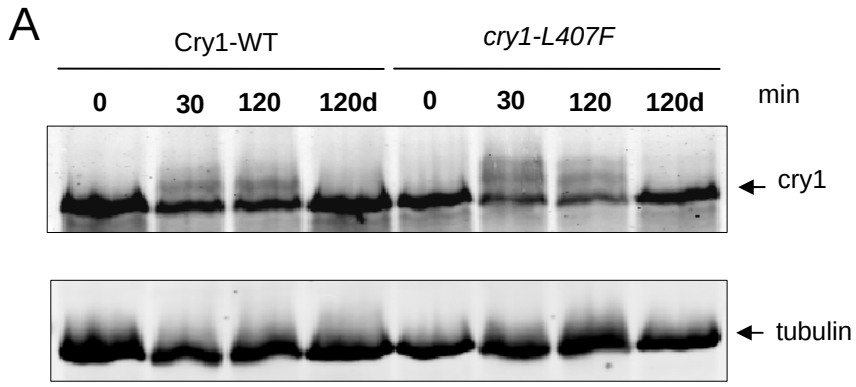
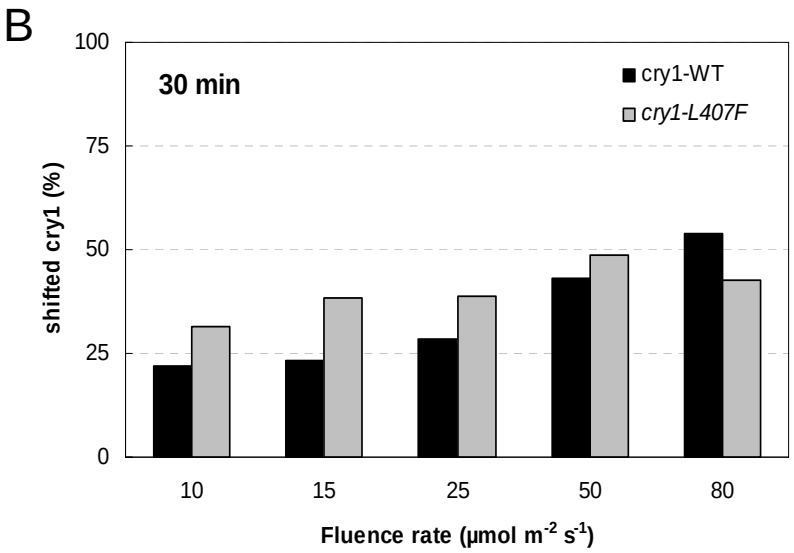

WT
cry1-L407F
Relative transcript levels
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
Time in red light (h)

Fig. 6


A
Cry1-WT
cry1-L407F
0
30
120
120d
0
30
120
120d
min
cry1
tubulin


B
30 min
cry1-WT
cry1-L407F
shifted cry1 (%)
100
75
50
25
0
10
15
25
50
80
Fluence rate (µmol m-2 s-1)

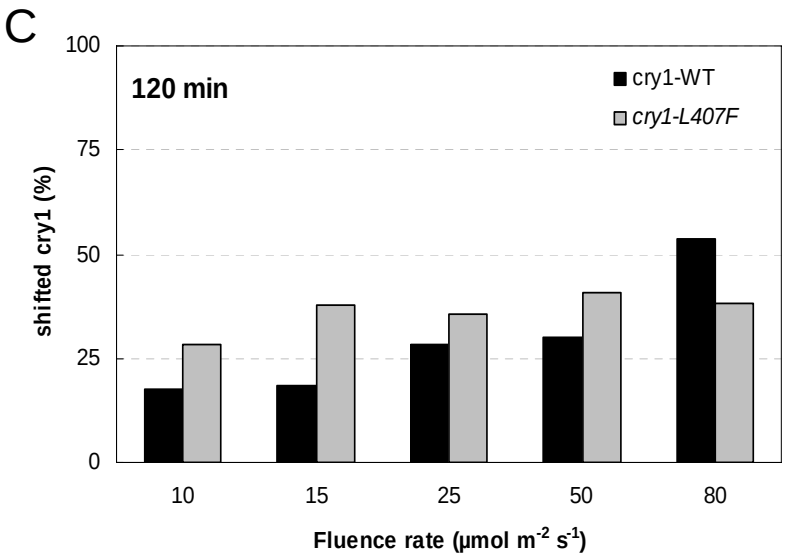

C
120 min
cry1-WT
cry1-L407F
shifted cry1 (%)
100
75
50
25
0
10
15
25
50
80
Fluence rate (µmol m-2 s-1)

Fig. 7

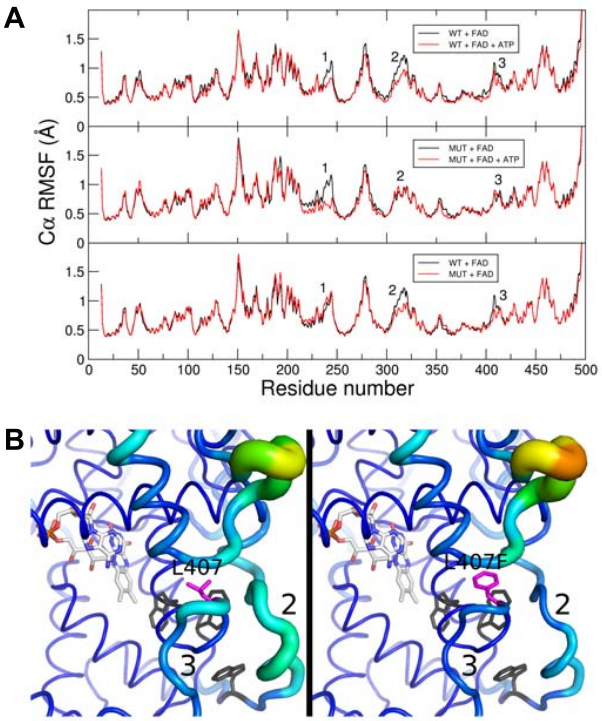

A
WT + FAD
WT + FAD + ATP
MUT + FAD
MUT + FAD + ATP
WT + FAD
MUT + FAD
Cα RMSF (Å)
Residue number
B
L407
L407F
2
3

Fig. 8

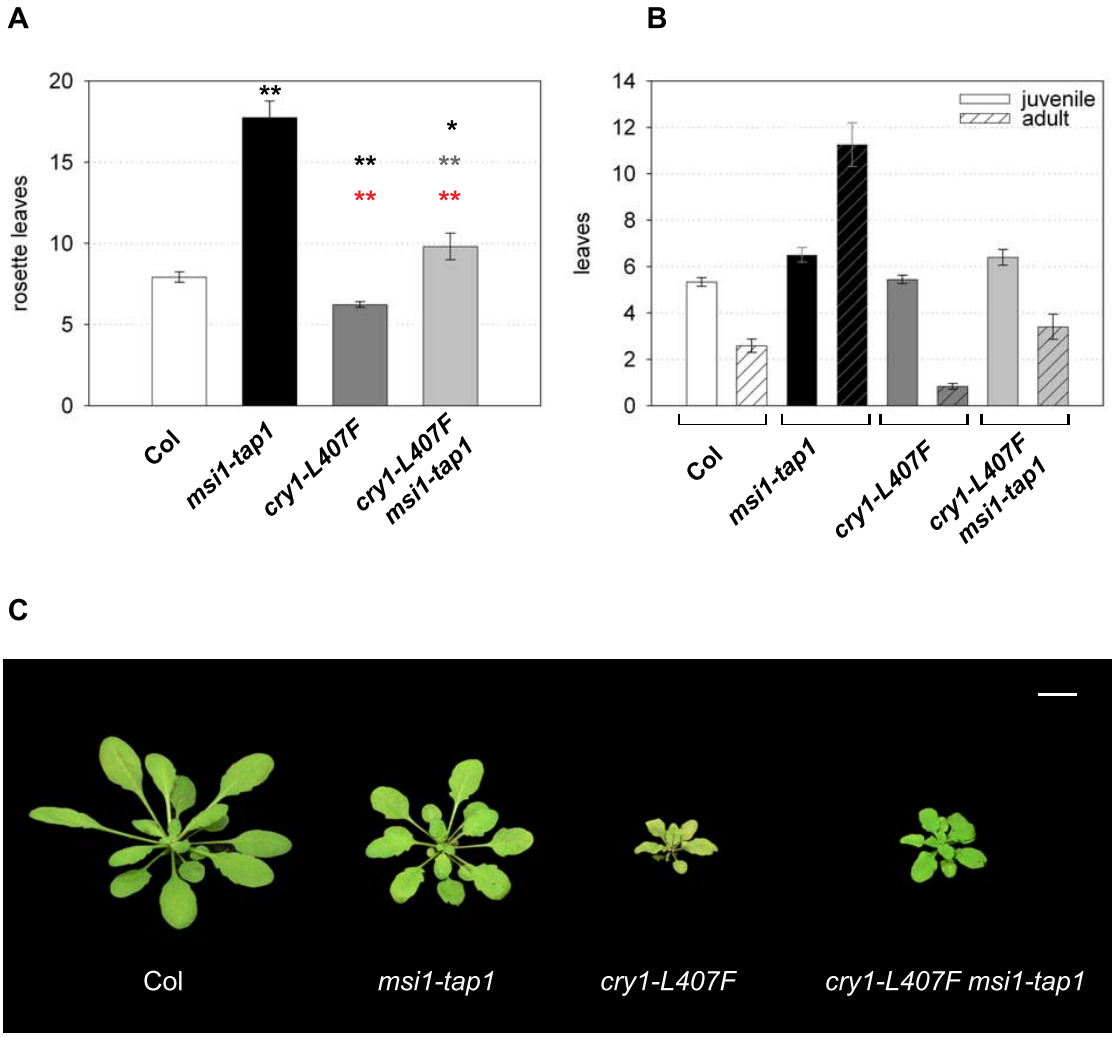

A
B
C
rosette leaves
leaves
juvenile
adult
Col
msi1-tap1
cry1-L407F
cry1-L407F msi1-tap1
Col
msi1-tap1
cry1-L407F
cry1-L407F msi1-tap1

Fig. S1

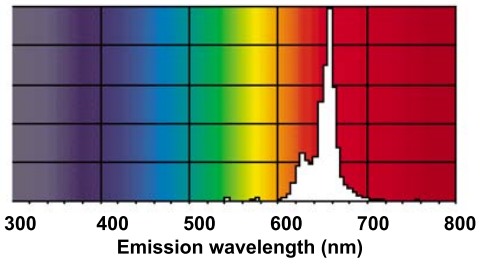


Fig. S2

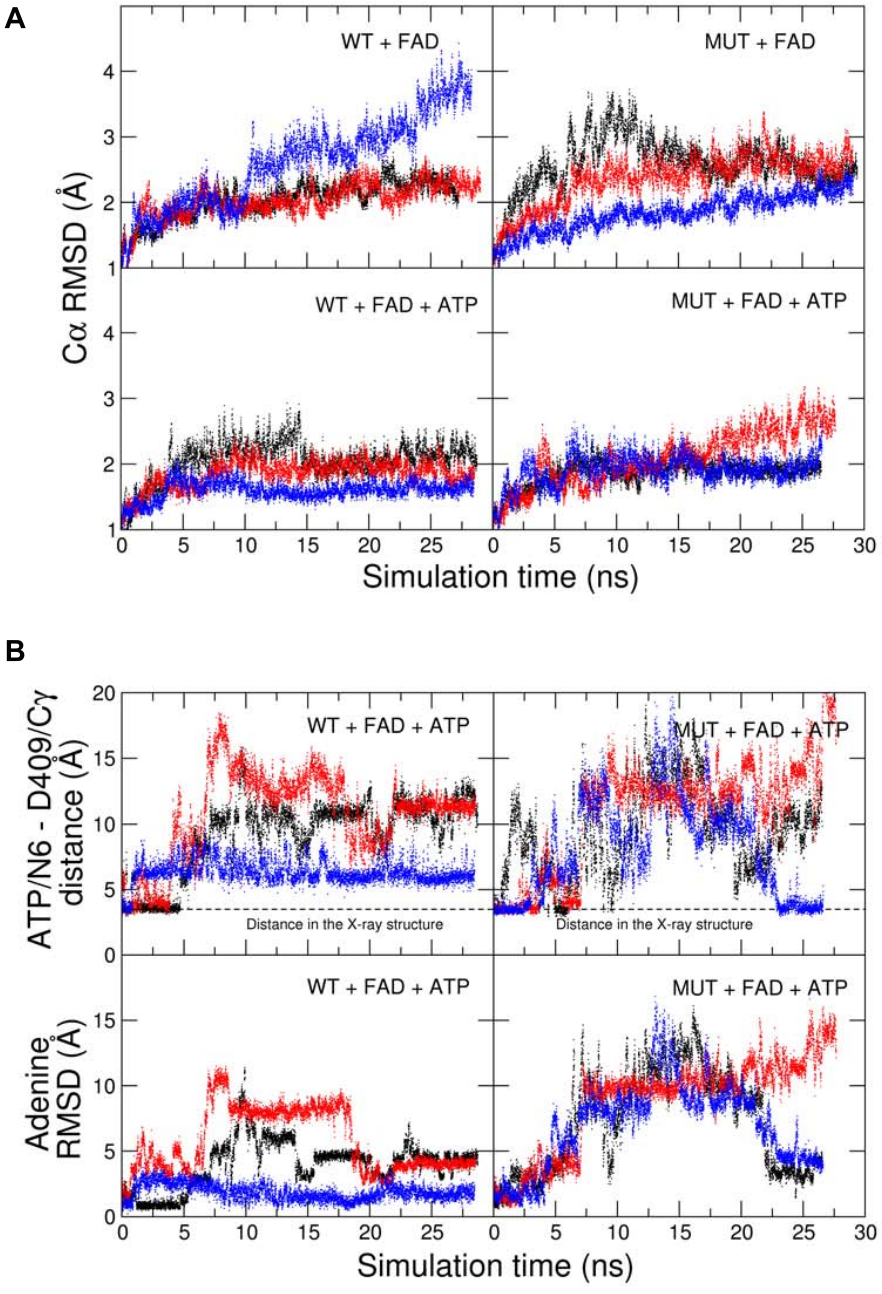

A
WT + FAD
MUT + FAD
WT + FAD + ATP
MUT + FAD + ATP
Cα RMSD (Å)
Simulation time (ns)
B
WT + FAD + ATP
MUT + FAD + ATP
Distance in the X-ray structure
ATP/N6 - D409/Cγ distance (Å)
Adenine RMSD (Å)
Simulation time (ns)

Fig. S3

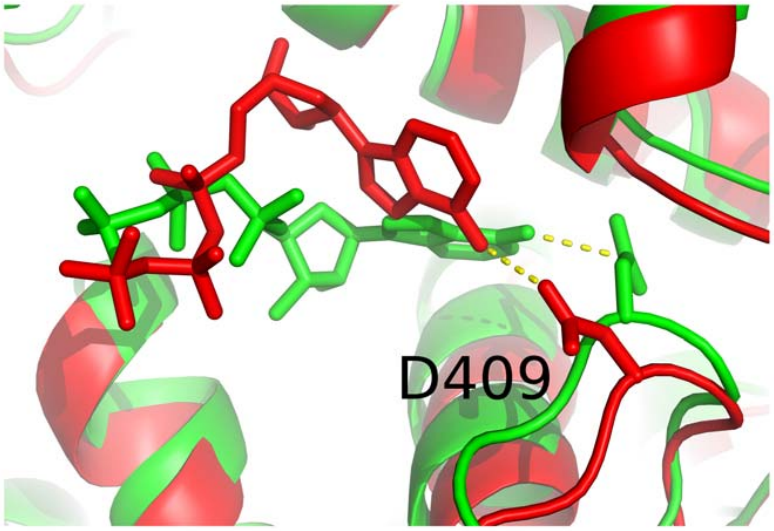


Fig. S4

Chapter 6

Computational optimization of the caps of a designed armadillo repeat protein

Pietro Alfarano et al.

Manuscript in preparation.

May 26, 2010

1 Introduction

Molecular recognition is a very important aspect of biology and it is involved in many biological processes. There are several classes of protein binders that recognize peptides: protein domains such as SH2, PTB, PDZ, SH3, and WW; the major histocompatibility complexes; antibodies; and repeat proteins. The latter is very interesting because of the modular, repetitive nature of the proteins.

Repeat proteins are made up by several tandem repeats of α-helical structural unit, which create an extended superhelical structure. The superhelical structure is very suited to generate an interface for the recognition of polypeptides in the extended conformation.

Armadillo repeat proteins are made up by 42 aminoacids-long repeats formed by three α-helices, named H1, H2, and H3. H3 helix is the longest and constitutes the binding surface for the polypeptide. Each repeat binds two peptides in a context of a longer peptide. The peptide bonds of peptide backbone is contacted by a conserved asparagine residue of each repeat, and other sidechains on the binding surface provide the specificity of binding. Figure 1 shows an importin α armadillo repeat protein (10 repeats) bound to a polypeptide. It is important to notice that internal repeats have a solvent accessible surface and two buried surfaces, where they contact neighboring flanking repeats. Only the first and the last repeat, called N- and C-terminal caps, have only one buried surface and they are shorter than the internal ones (see figure 2).

Parmeggiani and colleagues[12] designed artificial armadillo repeat proteins optimizing the hydrophobic core derived from a consensus sequence with computational methods. The consensus sequence was obtained by multiple alignment of single armadillo repeat modules to generate an unique stable internal module sequence. The computational optimization of the hydrophobic core was needed because the consensus sequence protein had molten globule properties. Interestingly, they found that the quality of the NMR spectra was pH dependent, probably due to ionizable side chains of lysine residues. Only after changing the acidity of the solution from pH=7 to pH=11, most of the peaks in the 2D-NMR spectrum appeared, although with low dispersion (see figure 3). Interestingly, the mutation of two lysines to glutammines in every internal repeats generated a mutant with better 2D-NMR spectrum at lower pH (see figure 4). Therefore

the pH-dependence of the quality of the NMR spectrum was attributed to the lysines, which have a side chain pK_a of 10.5 circa. A stable internal module repeat was needed because it can be randomized in positions which provide binding sites for amino acids side chains, and therefore screened for binding. The mutated internal repeat are very likely to be less stable, and for this reason a very stable internal repeat is needed.

The aim of this work was to improve the signal dispersion of NMR spectra by reducing the flexibility of the computationally optimized protein with the consensus sequence (YM_4A). Since mutating the sequence of internal repeats is experimentally time-demanding, we focussed on the optimization of the N- and C-terminal caps. The absence of a crystal structure of YM_4A prompted us to design a progressive optimization method of homology models based on consecutive cycles of molecular dynamics simulations and clustering analysis: to sample the conformations of the protein and to choose the most favourable conformations of repeat dimers, respectively. Subsequently, a new model is built from the clustered dimers taking advantage of the modular nature of the protein as described farther in the methods. Each optimization cycle produced a new generation of models having a previous generation model as input. The flexibility of the protein was assessed from the simulation trajectories by root mean square fluctuation (RMSF) analysis and by the calculation of the configurational entropy. The optimization procedure was applied until when the entropy difference between two consecutive generations was negligible.

Four mutations and a deletion, which in the simulations reduced the flexibility of the protein, were proposed and introduced in YM_4A. The mutant showed better 2D-NMR plots and also a better chemical and temperature denaturation stability.

2 Methods

The consensus armadillo sequence is called YM_4A elsewhere[12], Y refers to the sequence of the N-terminal cap; M_4 refers to the sequence of the four internal repeats; A refers to the sequence of the C-terminal cap. Here, the YM_4A sequence is called KK because of the presence of lysine at position 26 and 29 repeat-wise of the internal repeats, that are positions 60, 102, 144, 186 and 63, 105, 147, 189, respectively. The QQ mutant was derived from the KK model with aforementioned lysines mutated to glutamines. Further studied mutations are reported in table 2.

The following nomenclature is henceforth adopted: Ri stands for the i-th internal repeat; Ri-Rj stands for the dimer composed by the i-th and j-th repeats; Ncap and Ccap are the N-terminal and the C-terminal caps, respectively.

2.1 Molecular dynamics simulations

Langevin dynamics simulations were performed at 300K using the program CHARMM and the implicit solvent FACTS[5]. The protein was modeled according to the united atom CHARMM PARAM19 force field. To effectively compare simulations with experiments, the protein side chains were titrated at pH=7.4, that is, the sidechains of aspartates and glutamates were negatively charged, those of lysines and arginines were positively charged, histidines were considered neutral, the N-terminus was positively charged and the C-terminus negatively charged. SHAKE was applied to the hydrogens allowing an integration step of 2fs. Different initial random velocities were assigned to every simulation.

Each simulation consisted of a 0.2ns heating phase, 0.4ns equilibration phase and 30ns production phase, if not specified differently. About 10.5 hours on a core of a XEON 5410 Quadcore CPU running at 2.33GHz are required for a 1ns trajectory of the KK model (nearly 2220 atoms).

Explicit solvent molecular dynamics simulations were performed at 300K using the program CHARMM. The protein was modeled according to the all-hydrogen CHARMM force field (PARAM22 with CMAP correction)[9][8] and TIP3P water model[7]. To effectively compare simulations with experiments,

the protein sidechains were titrated at pH=7.4 (see above). The protein was inserted into a a water box where each atom of the protein had at least 13 Å distance from the boundary. Chloride and sodium ions were added to neutralize the total charge of the system at a concentration of 200mM. To avoid finite-size effects, periodic boundary conditions were applied. Different initial random velocities were assigned to every simulation. Long-range electrostatics effects were taken into account by the Particle Mesh Ewald summation method[3]. The temperature was kept constant by the Nosé-Hoover thermostat[11][6] while the pressure was held constant at 1 atm by applying the Langevin piston pressostat. SHAKE was applied to the hydrogens allowing an integration step of 2fs. Lookup tables[10] for the calculation of pair wise nonbonded interactions (van der Waals and Coulomb) were used to increase efficiency. Different initial random velocities were assigned to every simulation.

2.2 Clustering Analysis

A clustering analysis was applied to the snapshots of simulations trajectories to obtain the most populated conformers. Conformations were grouped together according to their RMSD (see farther) distance. At the end of the clustering procedure, several clusters were obtained and for every one the representative was defined. The cluster representative is the structure that differs the least from all the other conformations in the same cluster.

The first nanosecond of every production trajectory was removed from the analysis and one tenth of the snapshots, that is every 20ps, was analyzed. Conformational clustering was performed using the program "Cluster" version 3.2 by Michael Schaefer (Syngenta Crop Protection AG, unpublished work). We clustered the conformations of contiguous repeat dimers: the N-terminal cap and the first internal repeat (Ncap-R1); the last internal repeat and the C-terminal cap (R4-Ccap); and all the internal repeats (Rn-Rn+1). The conformations of the internal repeat dimers (R1-R2, R2-R3 and R3-R4) were collected together to generate a single model of the internal repeat dimer, therefore the dimers R1-R2, R2-R3 and R3-R4 are identical. All structures were clustered on all C_α atoms, except the first two residues for Ncap and the last of Ccap, and a selection of C_γ atoms. C_γ atoms were included to cluster the sidechains orientation

in the hydrophobic core. For this purpose, C_γ atoms of of lysine, glutammine, asparagine, glutamate and arginine were not included. A cutoff of 1.5 Å was chosen for clustering. This value, found by trial and error, is optimal to discern between non homologous conformers. The clusters representatives of the most populated cluster were used for modelling.

2.3 Trajectory analyses

Root mean square deviations and root mean square fluctuations of an atom j, RMSD and RMSF_j respectively, were calculated with CHARMM, and their formula is:

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x_{i,ref})^2 + (y_i - y_{i,ref})^2 + (z_i - z_{i,ref})^2}$$

$$\mathrm{RMSF}_j = \sqrt{\frac{1}{N_f}\sum_{i=1}^{N}(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2 + (z_i - \tilde{z}_i)^2}$$

where N is the number of atoms and N_f is the number of frames in the trajectory segment where RMSF is calculated; x_i, y_i, z_i are the coordinates of the atom i; $x_{i,ref}$, $y_{i,ref}$, $z_{i,ref}$ are the coordinates of the atom i in the reference structure; $\tilde{x}_i$, $\tilde{y}_i$, $\tilde{z}_i$ are the coordinates of the atom i in the average structure. The reference structure for RMSD analyses was the starting structure used in the dynamics. The average structures and RMSF were calculated for every 2ns segments of trajectory. For 30ns of simulation time, 15 values of RMSF were therefore calculated.

The quasiharmonic entropy was computed using the covariance matrix of the atomic fluctuations[1] using the program Wordom[13]. The quasiharmonic approximation was shown to yield an upper bound for the true entropy and is calculated with the following equation:

$$S_{\mathrm{ho}} = k\sum_{i}^{3n-6}\frac{\hbar\omega_i/kT}{e^{\hbar\omega_i/kT}-1} - ln(1 - e^{\hbar\omega_i/kT}),$$

where ω_i are the quasiharmonic frequencies obtained from the eigenvalues λ_i. λ_i are obtained from the diagonalized mass-weighted covariance matrix σ', which is obtained directly from MD simulations.

Global and local entropies were computed. Global entropies, calculated on all C_α, were divided by the number of residues to compare models of different lenghts (i.e. KK has 243 residues while $\Delta 5$ has 242 residues). Local entropies were calculated for a subset of atoms spanning all repeat dimers (i.e. Ncap-R1, R1-R2, R2-R3, R3-R4 and R4-Ccap).

2.4 Model generation

The first generation KK model was derived from three homology models built with Insight II (Accelrys Inc.) by mapping the YM_4A sequence on the crystallographic structure of three natural armadillo repeat proteins: yeast karyopherin (importin α), mouse importin α and murine β-catenin (PDB accession numbers: 1EE4, 1Q1T, 2BCT, respectively). A single implicit solvent MD simulation was run for each homology model, while, for further generations models, six molecular dynamics simulations were run.

The models were constructed assembling together the most probable structure of the three repeat dimers obtained from MD simulations through clustering analysis: 8706 conformations of the dimer Ncap-R1 (6879 for the homology models), 26118 conformations of the three internal repeat dimers (51579 for the homology models, because they have a different number of internal repeats, see table 1) and 8706 (6879 for the homology models) conformations of the R4-Ccap dimer is used for clustering. The resulting representatives of the most populated cluster of the dimers were assembled superimposing C_α atoms of the common repeat (see figure 6). For istance, the dimers Ncap-R1 and R1-R2 are superimposed on the common R1 monomer. The repeat R1 of the dimer R1-R2 is deleted to mantain the original interface between Ncap-R1. The procedure is repeated to build the whole model. Finally, the pair R3-R4 was deleted to mantain the interface between R4 and C. Missing peptide bonds between the assembled dimers were generated by CHARMM.

The optimization of the initial position of hydrogens and subsequent energy minimization were performed with the CHARMM[2] PARAM19[9] united atom force-field with distance-dependent dielectric function. Loops connecting α-helices were relaxed through four minimization cycles consisting of 100 iterations of steepest descent and 200 steps of conjugate gradient algorithms with gradually

decreasing harmonic constraints on the C_α atoms of the helices (k=10, 5, 1 and 0.1 Kcal/mol/Å^2). The system was further optimized with the presence of the implicit solvent FACTS without constraints by 100 steps of steepest descent and 200 iterations of conjugate gradient, followed by an adopted basis Newton-Raphson minimizer until an energy gradient of 0.02 Kcal/mol/Å^2 was reached.

The program WitnotP (Armin Widmer, Novartis Pharma, not published) was used for the modelling.

2.5 Geometry analysis

Interrepeat twist angle (see figure 18C) was determined as the angle between the axis of the helix 3 of two repeats (segment $\overline{\mathrm{NM}}$ in figure 18B). Helix 3 is the longest helix of a repeat. The axis of the helix has been calculated as the vector sum of the vectors defined by the carbon and oxygen atoms of the peptide bonds.

To determine the interrepeat bend angle, a plane on which the repeat lies had to be defined. For this purpose, three points were marked (see figure 18B): N and M, two point on the axis of the helix 3 and P, the center of gravity of helix 1 and helix 2. The interrepeat bend angle (see figure 18C) was determined as the angle between the surface normals (segment $\overline{\mathrm{NQ}}$ in figure 18B) to the planes of two repeats.

The C_α RMSD to the original crystal structures used for building the starting model (1EE4, 1Q1T, 2BCT) was calculated with the program Witnotp, fitting to the crystallographic reference structures with the command *fit structal*. The core of the algorithm is from M.Gerstein and M.Levitt[4]. The main difference is that, to obtain an initial alignment needed for the Gerstein-Levitt procedure, the *fit structal* command uses a sequence alignment algorithm in which the usual scoring matrix is replaced by a function which rewards the alignment of C_α pairs with similar local backbone geometry. After the Gerstein-Levitt alignment step, the list of aligned pairs is purged by removing pairs for which the distance between the C_α in the aligned structure exceeds a given threshold. The threshold used in this work was 3 Å. This procedure allows, for example, a better fit for structures with different loop orientations, because the variability of the loops will not negatively influence the superposition.

2.6 Protein biochemistry and production of ^{15}N-labeled proteins

Sorry, the protocols of the experiments are still missing!!!

Experimental conditions.

ANS binding: protein, 10 μM; ANS, 100 μM; buffer PBS 150, pH 7.4

CD spectra: protein, 10 μM; buffer PBS 150, pH 7.4

SEC: Superdex 75 column; buffer PBS 150, pH 7.4

Thermal denaturation: protein, 10 μM; buffer PBS 150, pH 7.4

Chemical denaturation: protein, 10 μM; buffer PBS 150, pH 7.4; samples incubated overnight at 4 C

The repeat proteins YM3A, YM4A and their respective cap mutants (see somewhere in Gauthams part) were expressed in the E. *coli* strain M15 in M9 minimal medium containing ^{15}N-NH_4Cl as the sole nitrogen source: 5 ml of overnight culture (LB medium, 1% glucose, 100 mg/l ampicillin and 25 mg/l kanamycin, 37 C) was used to inoculate 1 L cultures (M9 medium, 1% glucose, 150 μM thiamine, 30 mg/ml ampicillin and 25 mg/ml kanamycin, 37 C). At OD_{600} = 0.6 (after 6 to 8 hours), the cultures were induced with 1 mM IPTG and further incubated for 4 hours.

2.7 NMR Spectroscopy

NMR experiments were carried out using 250-500 μM solutions of YM3A, YM4A, and mutants in 50 mM phosphate buffer, 150 mM NaCl, pH 7.4. All NMR spectra were acquired at 310 K on Bruker Avance 600 or 700 MHz spectrometer equipped with cryoprobes. Experiments were selected from the Bruker standard pulse sequence library. All spectra were processed in TOPSPIN 2.1 and spectra were evaluated using the program Topspin 2.1.

3 Results

3.1 Initial homology models and KK models

Implicit solvent MD simulations were performed for every one of the homology models derived from the YM_4A consensus sequence (see table 1). RMSD plots of the C_α atoms show deviations of about 7 to 10 Å from the starting model, which indicate structural instability (see figure 8, top). Visual analysis of the trajectories suggested that the protein structure undergoes a rigid body rearrangement of the repeats. On the contrary, the simulations started from the crystal structure of yeast importin α (1EE5) with its wild-type sequence (see figure 8, middle) were structurally stable, being the RMSD less than 5 Å. These results provide evidence for the suitability of the simulation protocol and indicate that the fold of the homology models is not optimal.

Therefore, the first generation KK model for NR_4C was built from the homology models simulations as described in the methods part. The population of the clusters of the N-R1 and internal repeat dimers, whose representative structures were used for building the model, belongs uniquely to the mouse importin α simulation (1Q1T); while that of the dimer R4-C belongs uniquely to the murine β-catenin simulation (2BCT). Remarkably, the KK model simulations are more stable than those of the original homology model. The RMSD does not exceed 4 Å for six independent simulations (see figure 8, bottom).

Furthermore, three short molecular dynamics simulations of the KK model (6.5ns, 6.6ns and 4.9ns) were run in explicit solvent. Also in this case the fold is stable, being the RMSD around 4 Å(supplementary materials). Interestingly, in one simulation, a water molecule permeated the interface R4-C close to a buried glutamine (Q240) (see figure 9). In the crystal structure of β-catenin (2BCT), this position is occupied by a methionine (M662), which is also buried. The hydration of Q240 position was not noticed in implicit water simulations because the implicit solvent is not able to reproduce the granular properties of water. Nevertheless, the C-terminal cap showed more conformational instability than the internal repeats also in the implicit solvent simulations. In fact, the average RMSD of the N-terminal cap (2 Å) is higher than that of the internal repeat (1 Å) (supplementary materials).

A second generation KK model was built according to the clustering analysis based on simulations started from the first generation KK model, and six MD simulations were run. Third, fourth and fifth generation KK models were also built following the same protocol.

3.2 QQ models and mutational study

It was previously observed from NMR experiments [12] that the KK protein generates better resolved spectra only at pH higher than 10, while the QQ mutant provides an increase in spectra quality already at pH 8.

Therefore, a QQ model was derived from the second generation generation KK model and six molecular dynamics simulations were run. This model is called "second generation QQ model". Applying the same protocol, third, fourth and fifth generation QQ models were subsequently built. It's important to notice that QQ models are completely independent from KK models from the third generation (see figure 7).

The RMSF plot of the second generation KK model showed high flexibility in the N-terminal and the C-terminal caps (see figure 10). To reduce the flexibility of N-terminal cap, mutations on three residues were introduced in second generation QQ models (see figure 19). The V24R mutation was introduced to favour the interrepeat salt bridge with E64 and remove the solvent exposed V24; while the serine at position 27 was replaced by an arginine to match the same position in the internal repeats. Finally, R32 was deleted because it showed high flexibility and to shorten the loop connecting the N-terminal cap and the first repeat.

The following mutations of the last helix of the C-terminal cap were also introduced: the buried Q240 was mutated into a leucine and the solvent exposed F241 was mutated into glutamine. Those two mutations were suggested by the aforementioned water molecule permeation and because of the high RMSF of this repeat. For this project, mutations of the internal repeats were not analyzed because of the higher synthetic feasibility of introducing mutations in the caps.

Third generation models of these mutants were produced since mutations were introduced on the third generation QQ model. Rounds of optimization were run up to the fifth generation. The complete list of mutations is reported

in table 2.

To assess the effects of the mutations on the whole protein flexibility, the quasiharmonic entropy of the fifth generation models was estimated (see figure 11). The quasiharmonic entropy is a measure of the configuration entropy. A reduced value of this quantity corresponds to a reduction of the structure flexibility and thus to an increase of fold stability. The plot shows that the average value of the entropy of the QQ model is lower than that of the KK model, but the decrease could not be significant because it lies within the error. The $\Delta 5$ model possesses a significant lower conformational entropy than any other mutations, in accord with the 2D-NMR spectra. To investigate the local effect of the mutations, the quasiharmonic entropy of the repeat dimers Ncap-R1, R1-R2, R2-R3, R3-R4 and R4-Ccap was calculated (see figure 12). The trend of QQ and $\Delta 5$ found for the total entropy is reproduced: the quasiharmonic entropy of $\Delta 5$ is lower than QQ for all the repeat pairs. Interestingly, the results of the N-cap and of Q240L mutants simulations are suggesting that the contribution to the entropy reduction is mainly located in the repeat dimer that the mutations affect, without affecting the internal repeats. Only for $\Delta 5$, where both caps are mutated, there is an overall decrease of whole and partial entropy.

The same conclusions can also be drawn from the analysis of the RMSF differences plot calculated on fifth generation models (see figure 13). In this plot, the KK model RMSF are used as a baseline to calculate the difference with respect to the other models. The red lines on the plot represent the position of the residues for the KK to QQ mutation (see table 2); the green ones the position of the residues mutated in the N-terminal cap; and the cyan and the orange lines show the mutations at the C-terminal cap. Mutations in the N-terminal cap (green lines) and in the C-terminal cap (cyan and orange lines) reduce the flexibility of the backbone around the mutated aminoacids. Moreover, the QQ mutation reduces the flexibility of the N-cap even if no mutation is introduced in this segment. The effect is reproducible, the QQ and Q240L models, which have no mutations in the N-cap, show a very similar flexibility reduction with respect to the KK model. The N-cap and $\Delta 5$ models, with the same mutations in the N-cap, show a very similar flexibility reduction, which is higher than for the QQ and Q240L models. These observations support the robustness of the

method.

3.3 Explicit water simulations

To further study and validate the results of the implicit solvent simulations, three long, independent explicit solvent MD simulations were run for each one of the KK, QQ and Δ5 models. The starting conformations of the simulations were the most visited conformations of the implicit solvent simulations, obtained by clustering on all the C_α atoms. The same analyses carried out on the implicit solvent simulations were performed on more than 80 ns of simulation time for every simulation. Therefore, implicit solvent and explicit solvent simulations are compared and analogies and differences are shown.

To assess the conformational flexibility, global and local entropy were calculated. Similarly to the implicit water simulations, the global entropy plot (see figure 14) shows that Δ5 is less flexible than QQ and KK. The partial entropy plot trend (see figure 15) is similar to the one of the implicit solvent simulations (see figure 15), but it shows that the average conformational flexibility of the KK model simulations in the Ncap-R1 dimer is lower than Δ5. This result is in disagreement with the implicit solvent simulations, where the flexibility of the Ncap-R1 dimer of KK is higher than the one of Δ5, and this discrepancy could be due to the reduced sampling (langevin dynamics vs. explicit solvent) and the reduced number of run simulations (6 for the implicit solvend and 3 for the explicit solvent simulations). For the other repeat dimers, KK is more flexible than QQ, and QQ is more flexible than Δ5, and this is in agreement with the implicit solvent simulations.

The RMSF difference plot (see figure 16) shows a trend which is similar to the implicit solvent simulations, but, in contrast, in the QQ and Δ5 model simulations, an increase of flexibility in the Ncap-R1 dimer with respect to the KK model is observed. The increase with respect to the KK model simulations is mainly involving the Ncap only (residues 23 to 33), where the mutations of the Δ5 model were introduced, and not the R1. Therefore, the mutations introduced in the Ncap (green lines in figure 16) and Δ5 reduce the flexibility of the aforementioned sequence segment in the implicit solvent simulations, while they increase the flexibility in the explicit solvent simulations. Interestingly, an

increase of flexibility in this segment is also observed in the QQ model, which has the same sequence of the KK model in this segment, and therefore its flexibility should be very similar to KK. One possible explanation is that the lysine to glutammine mutations of the internal repeats (see table 2)introduced in the QQ and $\Delta 5$ models can influence the N-cap. In fact, the mutated positions 60 and 63 are close to the N-cap (see figure 19) and the electrostatic effects due to the loss of the positive charge can be simulated better with the electrostatic treatment of an explicit solvent simulation (more distant cut-offs and Particle Mesh Ewald) than with an implicit solvent simulation (shorted cut-offs and no Particle Mesh Ewald). Moreover, the overall profiles of QQ and $\Delta 5$ are very similar, supporting the robustness of the method. In agreement with the implicit solvent simulations, the flexibility of the last residues (mutations in the Ccap) is lower for $\Delta 5$.

The interrepeat salt bridge stability (see figure 17) between the Ncap and the first repeat, that is between residues R27 (R24 in $\Delta 5$) and E64 (E63 in $\Delta 5$), respectively, is different between the KK, QQ and $\Delta 5$ models. In the fifth generation implicit solvent simulations (red curves in figure 17), KK and QQ simulations have a simular broad arginine-glutammate distance distribution, probably due to at least two superimposing populations, while $\Delta 5$ simulations are more unimodal. The same difference can be more clearly observed in the explicit solvent simulations (black lines in figure 17). In the explicit solvent simulations, the effect of the lysine to glutammine mutations introduced in the internal repeats of the QQ model has an effect on the salt bridge distribution, which is not observed in the implicit solvent simulations, where the distributions of the KK and QQ simulations are practically the same. This could be due to the aforementioned better treatment of electrostatics of explicit solvent simulations.

3.4 Geometrical analysis of YM_4A structure

To analyze the dynamical structural behavior of the three models, two geometry analyzes were performed. First, the inter-repeat bend and twist angles (see figure 18) were measured and compared to the ones of the three X-ray structures used for the generation of the three starting homology models (1EE4, 1Q1T, 2BCT). Second, a RMSD comparison between the conformations of the

simulations of the models and the three aforementioned X-ray structures was performed.

The geometry analysis of the explicit solvent simulations showed that KK, QQ and $\Delta 5$ models assumed conformations close to the ones of the X-ray structures used for building the original homology models. In fact, the interrepeat bend and twist angles in the range of those assumed in the X-ray structures (see figures 29-37 in the supplementary materials). RMSD comparisons between simulations and X-ray structures (see figures 38-46 in the supplementary materials and table 4) show that all the models are close to 1EE4 and 1Q1T (yeast and mouse importins, respectively), and none is close to 2BCT (murine β-catenin), if the superposition is carried out on all four internal repeats, Only if the superposition is carried out on three internal repeats, a low similarity similarity emerges to 2BCT, but this criterion is less stringent. This finding is in agreement with the higher similarity of the sequence of YM_4A internal repeats to importins rather than to β-catenins.

3.5 Biochemistry and molecular biology

Circular dichroism experiments showed that YM_4A QQ and relative mutants have helical secondary structures (see figure 20).

Size-exclusion chromatography showed that the original YM_4A consensus sequence and the mutants introduced in this work are monomeric species (see figure 21). The eluition volume of $\Delta 5$ is very close to QQ, and all the other mutants eluite at the same volume of QQ. The monomeric nature of the proteins was confirmed by MALS (data not shown).

YM_4A $\Delta 5$ has less solvent exposed hydrophobic surface than YM_4A QQ and other mutants according to the ANS binding (see figure 22).

Thermal and chemical denaturation experiments showed that YM_4A $\Delta 5$ is more resistant towards denaturation than other proteins (see figure 23 and 24).

3.6 Stability evaluation of YM_3A, YM_4A and their cap mutants using NMR spectroscopy

In order to evaluate the influence of different mutations or combinations of mutations in the capping repeats of YM_3A QQ and YM_4A QQ with respect to conformational rigidity and oligomerization, 1D ^{1}H-NMR spectra of all proteins were recorded (data not shown). Proteins were ranked according to signal dispersion in the amide- and methyl-region as well as the line width of their proton resonances. A subset of these, namely the original consensus proteins YM_3A QQ and YM_4A QQ, and three promising mutants thereof (see table 3), which appeared to be well structured, were expressed in their uniformly ^{15}N-labeled form and analyzed using [^{1}H,^{15}N]-HSQC NMR spectra.

The repetitive nature of the sequence and the inherently reduced signal dispersion within purely α-helical proteins is expected to result in rather poor signal dispersion. Accordingly, not all peaks in the [^{1}H, ^{15}N]-HSQC spectra were individually resolved (see figure 25 and 26). Nevertheless, signal dispersion is remarkably good and significantly further improves in the cap mutants. Interestingly, the effects due to the C-cap mutations Q240L and F241Q are stronger than those of the N-terminal V24R and R27S mutations and the deletion of R32. The combination of N- and Ccap mutations shows a synergistic effect resulting in the best signal dispersion for YM3A QQ V24R R27S ΔR32 Q240L F241Q and the respective YM4A QQ construct compared to other constructs of the same size. Due to overlap of peaks less than the expected number of peaks was usually observed, e.g. for YM_3A QQ xxx out of the expected zzz cross-peaks, for YM_3A QQ V24R R27S delR32 xxx (yyy expected), for YM_3A QQ Q240L F241Q xxx (yyy expected), and for YM_3A QQ V24R R27S delR32 Q240L F241Q xxx (yyy expected). It should be noted that the [^{1}H,^{15}N,]-HSQC spectra of YM_3A QQ and YM_4A QQ display reduced signal dispersion such that the indicated cross-peak numbers are only estimates. The line widths indicate that all proteins are monomeric species, in agreement with results obtained by size-exclusion chromatography (see figureo 21).

4 Discussion and conclusion

The synergistic approach, consisting of coupling static modeling and molecular dynamics simulations, allowed not only the generation and refinement of a plausible structure of the YM_4A armadillo repeat protein, but also the analysis and optimization of its dynamical properties.

Predictions of RMSF and conformational entropy analyses of mutants which show a decrease of flexibility were confirmed by experiments.

NMR measurements showed that the newly designed N- and C-caps mutants are significantly more stable and better structured than the corresponding initial constructs YM_3A QQ and YM_4A QQ. This result is in good agreement with predictions from molecular dynamics calculations and results from thermal and chemical unfolding experiments followed by CD spectroscopy, the ANS-binding behavior of the proteins and gel-filtration experiments.

The reduction of flexibility can not directly explain the increased resistance towards thermal and chemical denaturation. A possible explanation is that the solvent exposed hydrophobic surface of the protein is involved in the denaturation process, therefore a reduction of this hydrophobic surface, as measured by ANS binding, due to the mutations, which is most pronounced the $\Delta 5$ mutant, is beneficial.

5 Figures

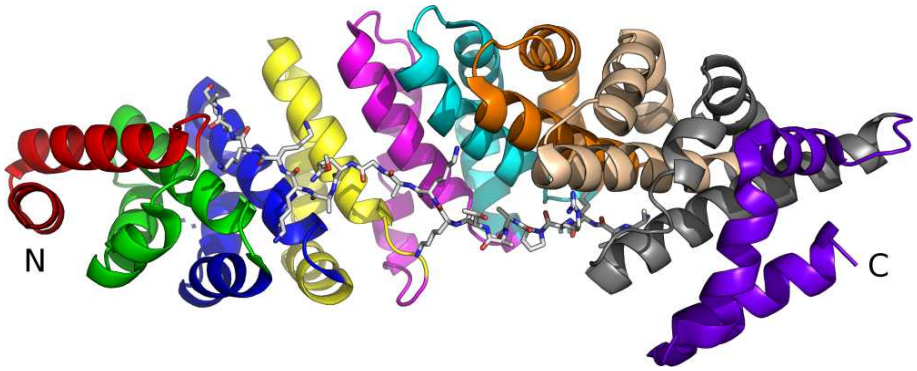


Figure 1: **An armadillo repeat protein bound to a peptide.** Importin α (pdb accession code: 1EE5) in complex with a nucleoplasmin NLS peptide is shown. Every repeat is colored differently and the NLS peptide is in sticks representation.

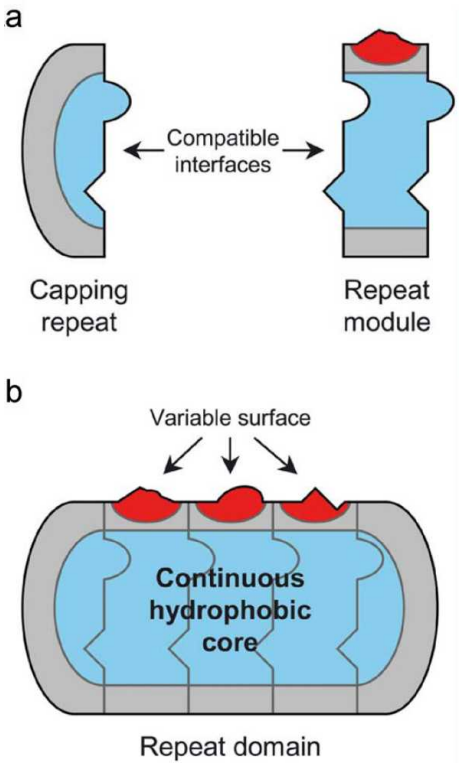


Figure 2: **Schematic representation of designed repeat proteins**. (a) Compatible interfaces allow the stacking of repeat modules. (b) The repeat modules form a continuous hydrophobic core, which is protected on both sides by capping repeats. In red is represented the variable surface responsible for binding and that is eventually randomized in a library. The hydrophobic core is shown in blue and the polar surface of the protein in gray. Reproduced from the doctoral thesis of Fabio Parmeggiani.

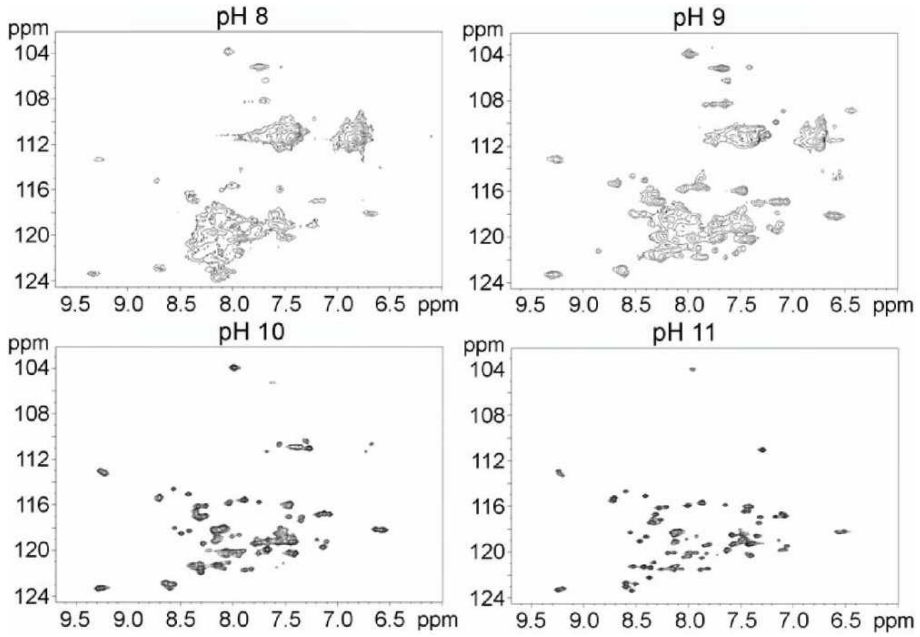


Figure 3: **The quality of the 2D-NMR spectra of YM_4A increases in a more basic environment .** Many peaks of the 2D-NMR peaks of appear between pH 8 and 9, although with low dispersion.

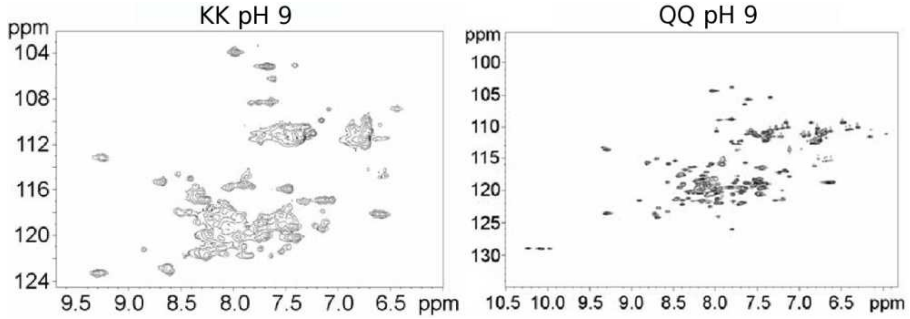


Figure 4: **YM_4A QQ has a better 2D-NMR spectrum than YM_4A KK at the same pH.** The mutation of two lysines into glutammines in every internal repeat of YM_4A improves the dispersion and intensity of the peaks in the 2D-NMR spectrum.

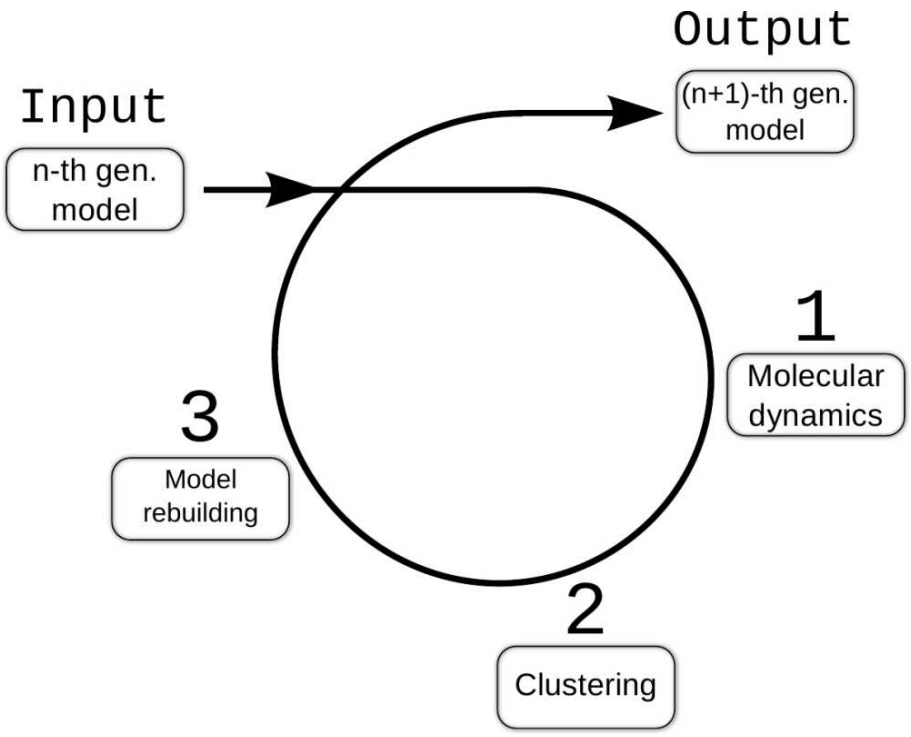


Figure 5: **Model generation procedure.** The optimization proceeds in three steps having the n-th generation model as input and producing a (n+1)-th generation model. Step 1: starting from the input model, six independent MD simulations in an implicit solvent are run to sample the repeat dimers conformations; step 2: snapshots are extracted from the MD trajectories and decomposed in adjacent repeat dimers. Dimers are then clustered to find the most populated conformation; step 3: the most populated conformations are used to build a new model (see figure 6).

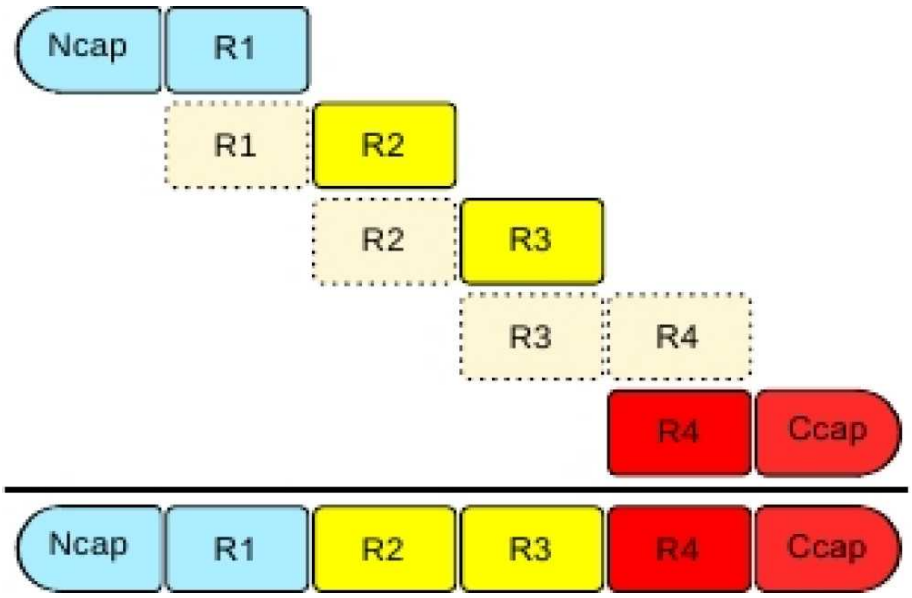


Figure 6: **Superposition procedure.** After the representatives of the most populated clusters are derived for the N-terminal cap and first internal repeat dimer (Ncap-R1), for the three internal repeat dimers (R1-R2, R2-R3, R3-R4) and for the fourth repeat and C-terminal cap dimer (R4-Ccap), these are superimposed as shown in the scheme. For istance, the dimers Ncap-R1 and R1-R2 are superimposed through the common R1 repeat. Afterwards, the repeat R1 of the dimer R1-R2 is deleted to mantain the original interface between Ncap-R1, as shown by dashed lines. The procedure is repeated to build the whole model. Finally, the pair R3-R4 is deleted to mantain the interface between R4 and Ccap.

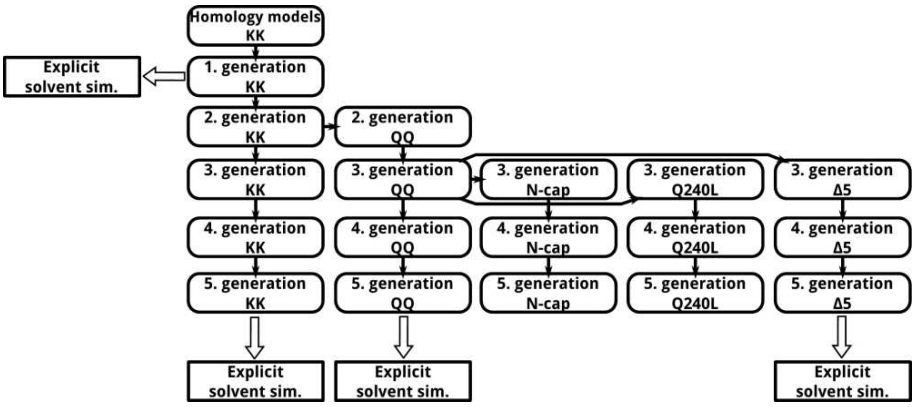


Figure 7: **Overview of all simulations.** Rectangles with round edges represent implicit solvent simulations, while sharp edges rectangles represent explicit solvent simulations. Small black horizontal arrows between rectangle pairs indicate models derived by mutations; small vertical black arrows represent the clustering procedure used to derive the starting conformation of the next-generation model. For instance, the second generation QQ model is derived from the second generation KK model upon mutation of the corresponding side chains.

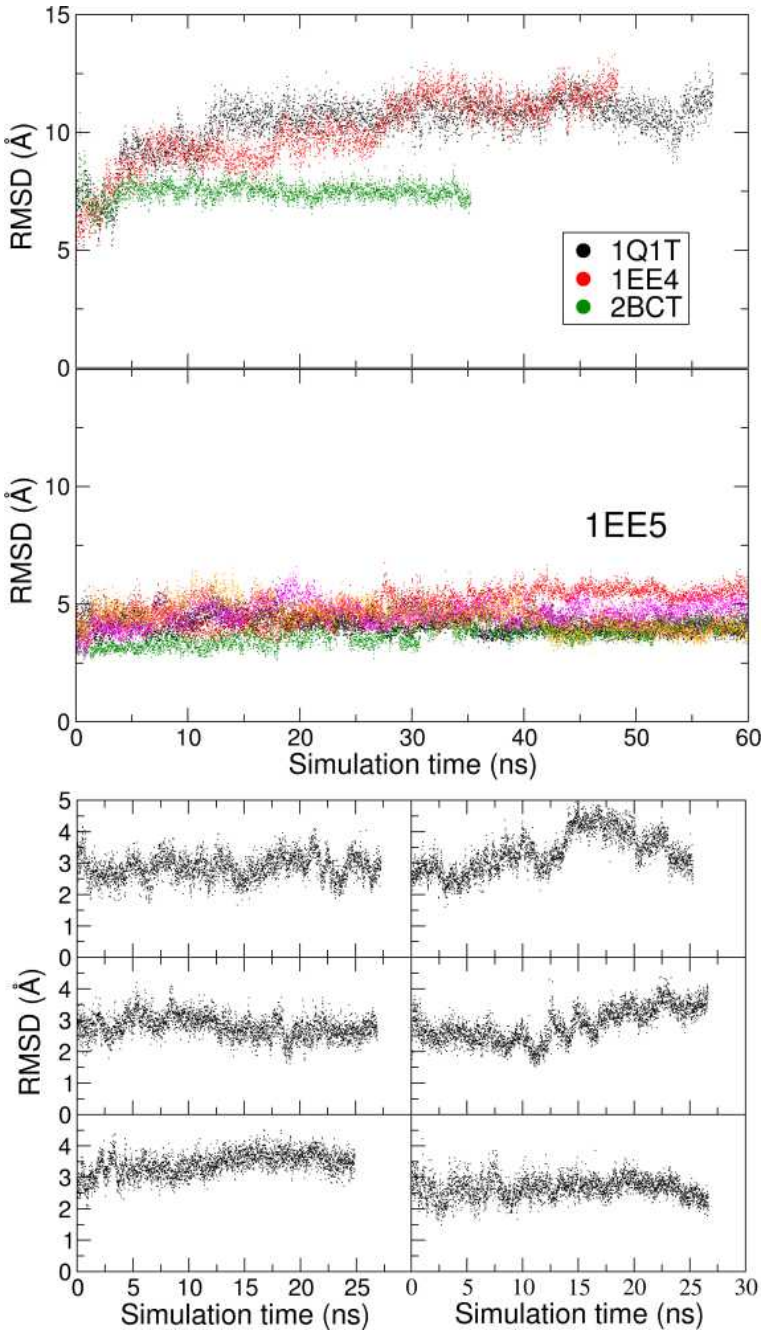


Figure 8: **Instability of homology models.** Top, implicit solvent MD runs started from homology models show an increase of RMSD, which indicates structural instability. In contrast, six MD simulations started from the crystal structure of the yeast importin α (1EE5, middle) and six MD simulations started from the first generation KK model (bottom) are stable.

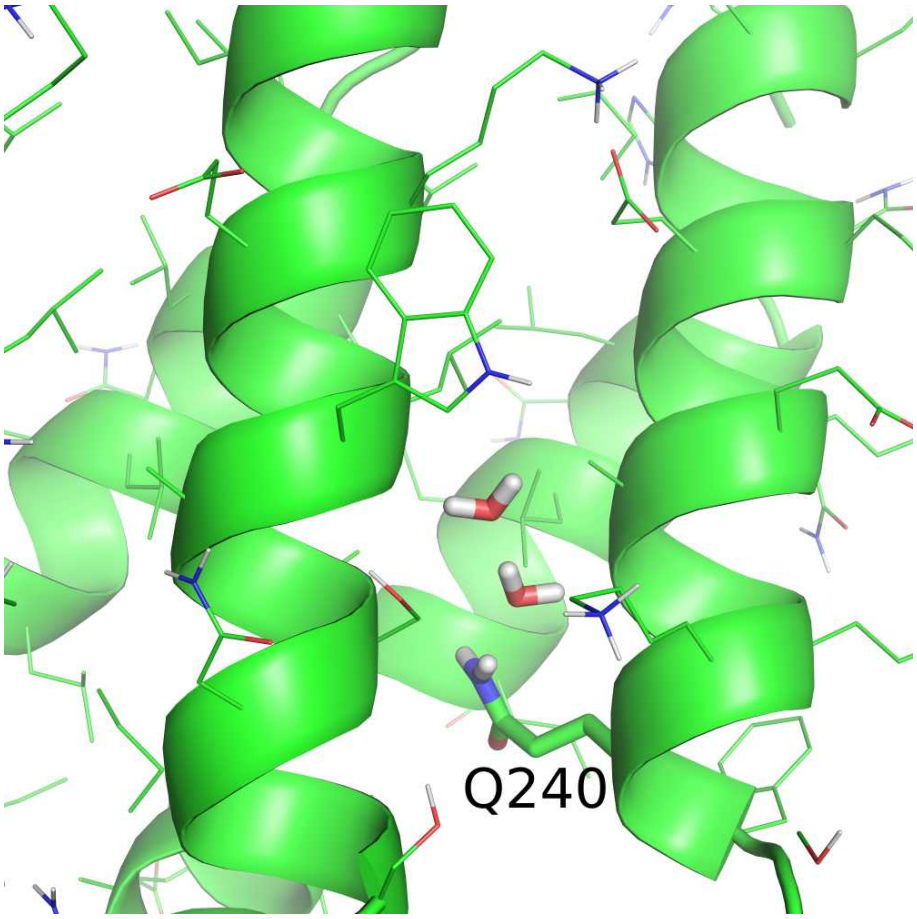


Figure 9: **Water molecules permeate into the R4-C interface.** In one explicit water simulation of first generation KK model, water molecules permeate into the hydrophobic surface between the fourth internal repeat and the C-terminal cap, close to buried Q240.

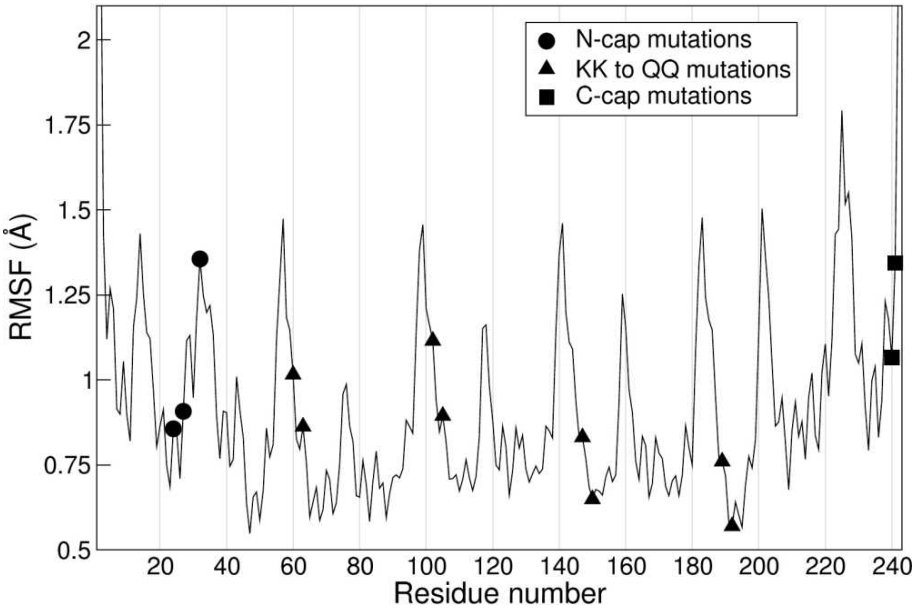


Figure 10: **RMSF plot of second generation KK simulations.** The circles show the position of the mutated residues V24R, R27S and ΔR32 in the Ncap; the triangles the internal repeat positions affected by the mutation K26Q and K29Q; the squares the mutations Q240L and F241Q in the Ccap. The mutation labels are explained in table 2.

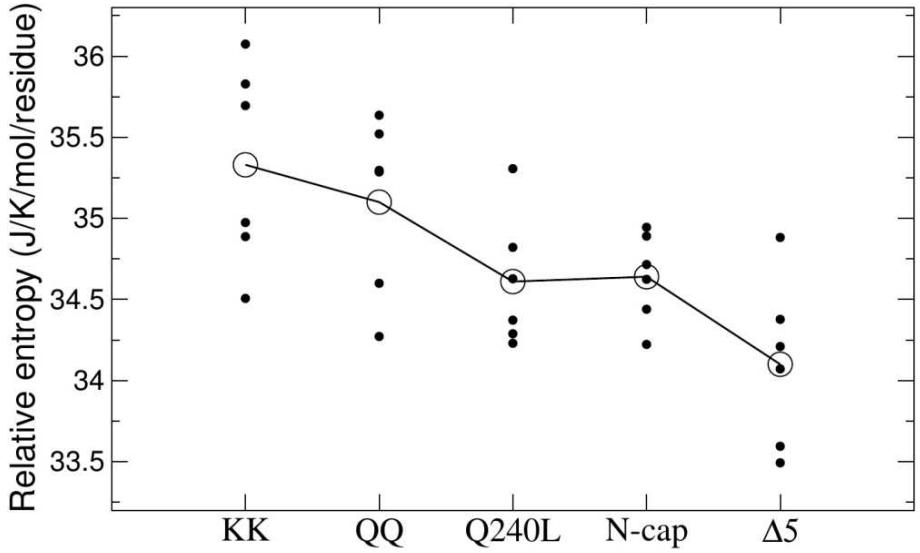


Figure 11: **Quasi harmonic entropy of fifth generation models.** The small filled circles are the individual calculations and the bigger empty circles are the averages. The entropy values have been divided by the number of residues to effectively compare the models with the deletion ΔR32 in the Ncap (Δ5 and Ncap). The mutation labels are explained in table 2.

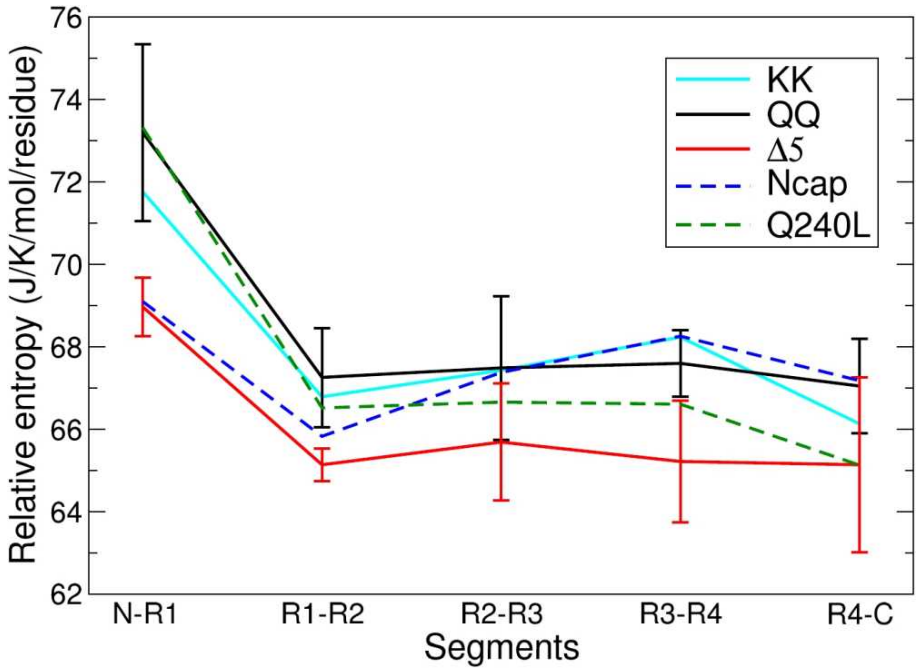


Figure 12: **Effect of the mutations on the quasiharmonic entropy of repeat dimers of fifth generation simulations.** The error bar is the standard deviation. Only the error bars of QQ and Δ5 simulations are shown. The entropy values have been divided by the number of residues in the segments to effectively compare the models with the deletion ΔR32 in the Ncap (Δ5 and Ncap). The mutation labels are explained in table 2.

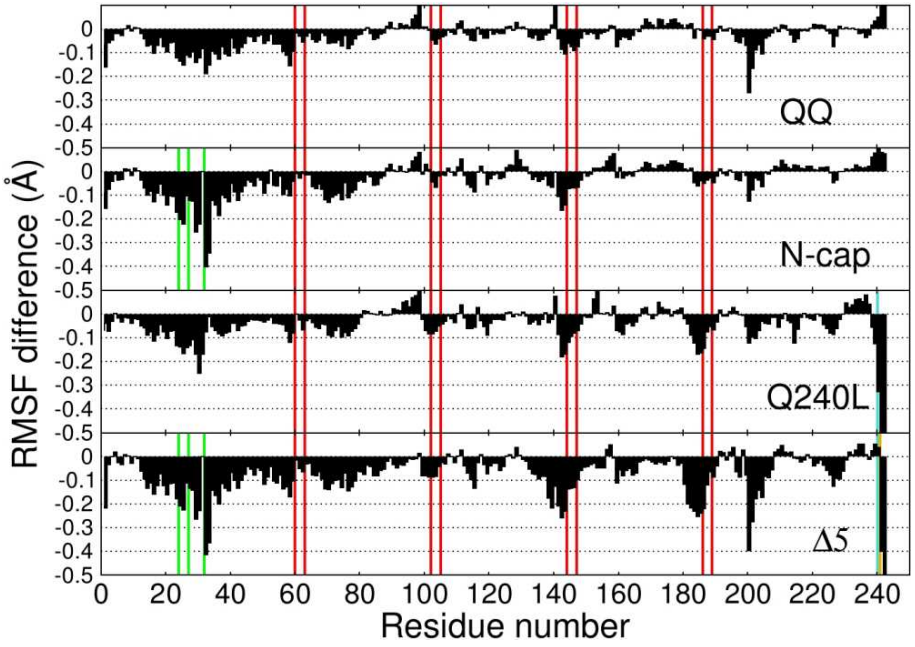


Figure 13: **RMSF comparison of fifth generation models simulations.** The topmost plot is the RMSF plot of the KK model. For every mutant, the RMSF difference to the KK model is plotted. Negative values indicate lower fluctuations with respect to the KK model. The lysine to glutammine mutations introduced in the QQ sequence are indicated by red lines; mutations at the N-terminal cap are indicated by green lines; the cyan and the orange lines the Q240L and F240Q mutation, respectively. The mutation labels are explained in table 2.

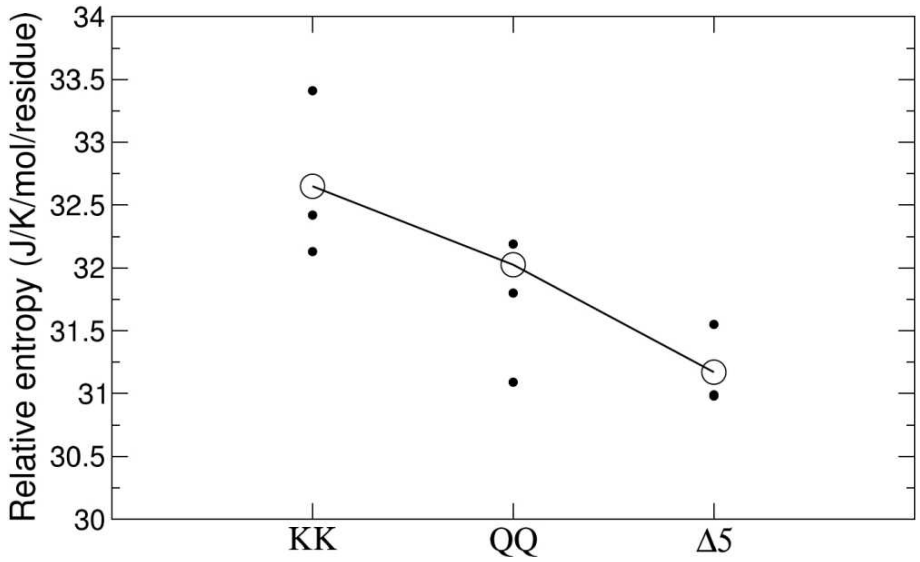


Figure 14: **Quasi harmonic entropy of explicit water simulations.** Three explicit water simulations were run per model. The small filled circles are the individual calculations and the bigger empty circles are the averages. The entropy values have been divided by the number of residues to effectively compare the $\Delta 5$ one (deletion ΔR32 in the Ncap). The mutation labels are explained in table 2.

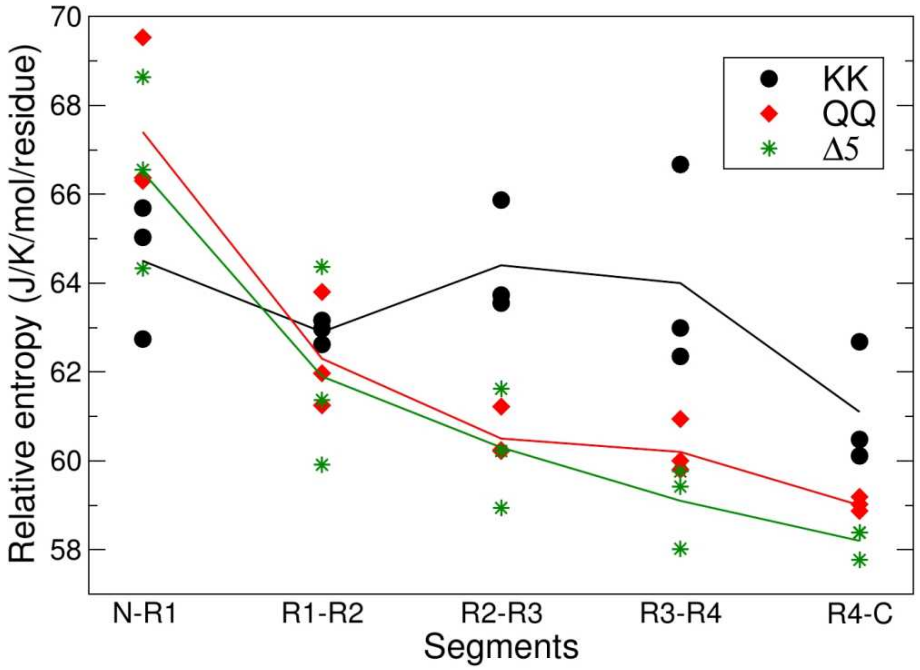


Figure 15: **Effect of the mutations on the quasiharmonic entropy of repeat dimers of explicit water simulations.** Three explicit water simulations were run per model. The entropy values have been divided by the number of residues in the segments to effectively compare the models with the deletion ΔR32 in the Ncap (Δ5 and Ncap). The mutation labels are explained in table 2.

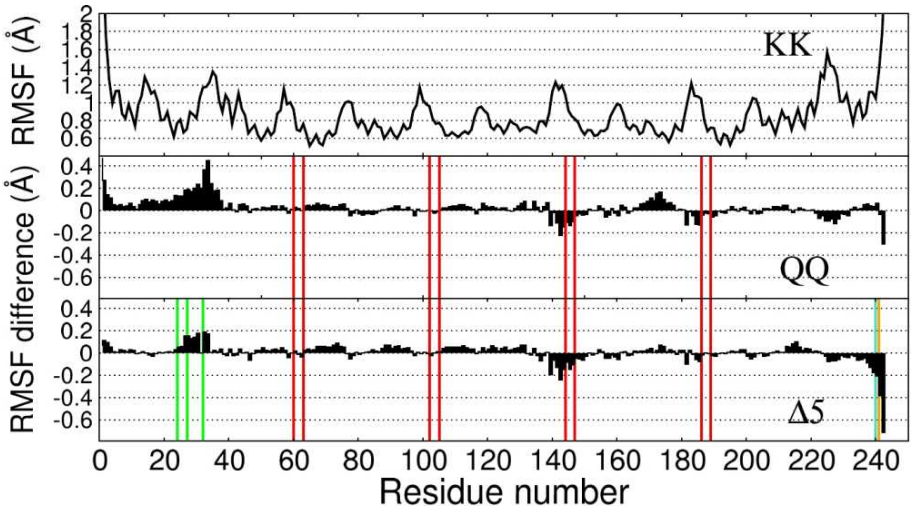


Figure 16: **RMSF comparison of explicit water models simulations.** The topmost plot is the RMSF plot of the KK model. For every mutant, the RMSF difference to the KK model is plotted. Negative values indicate lower fluctuations with respect to the KK model. The lysine to glutammine mutations introduced in the QQ sequence are indicated by red lines; mutations at the N-terminal cap are indicated by green lines; the cyan and the orange lines the Q240L and F240Q mutation, respectively. The mutation labels are explained in table 2.

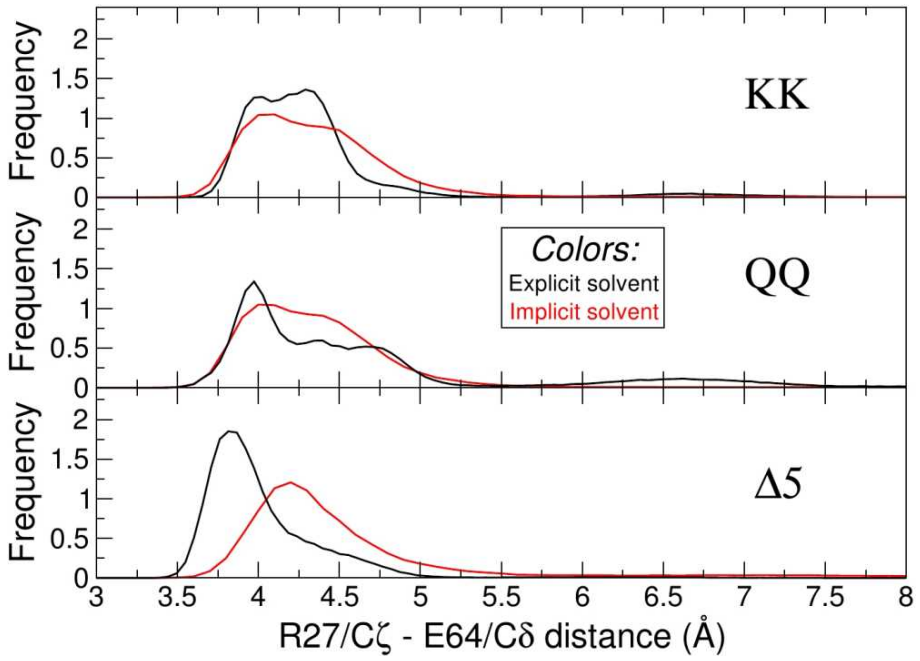


Figure 17: **Histogram of the salt bridge between the N-cap and the first repeat.** Histograms of the salt bridge between the N-terminal cap (R27, R24 in Δ5) and the first repeat (E64, E63 in Δ5). The black line refers to the explicit water simulations and the red one to the implicit solvent simulations. Shifting the arginine from position 27 to position 24 in the N-terminal cap, improves the unimodality of the salt bridge population.

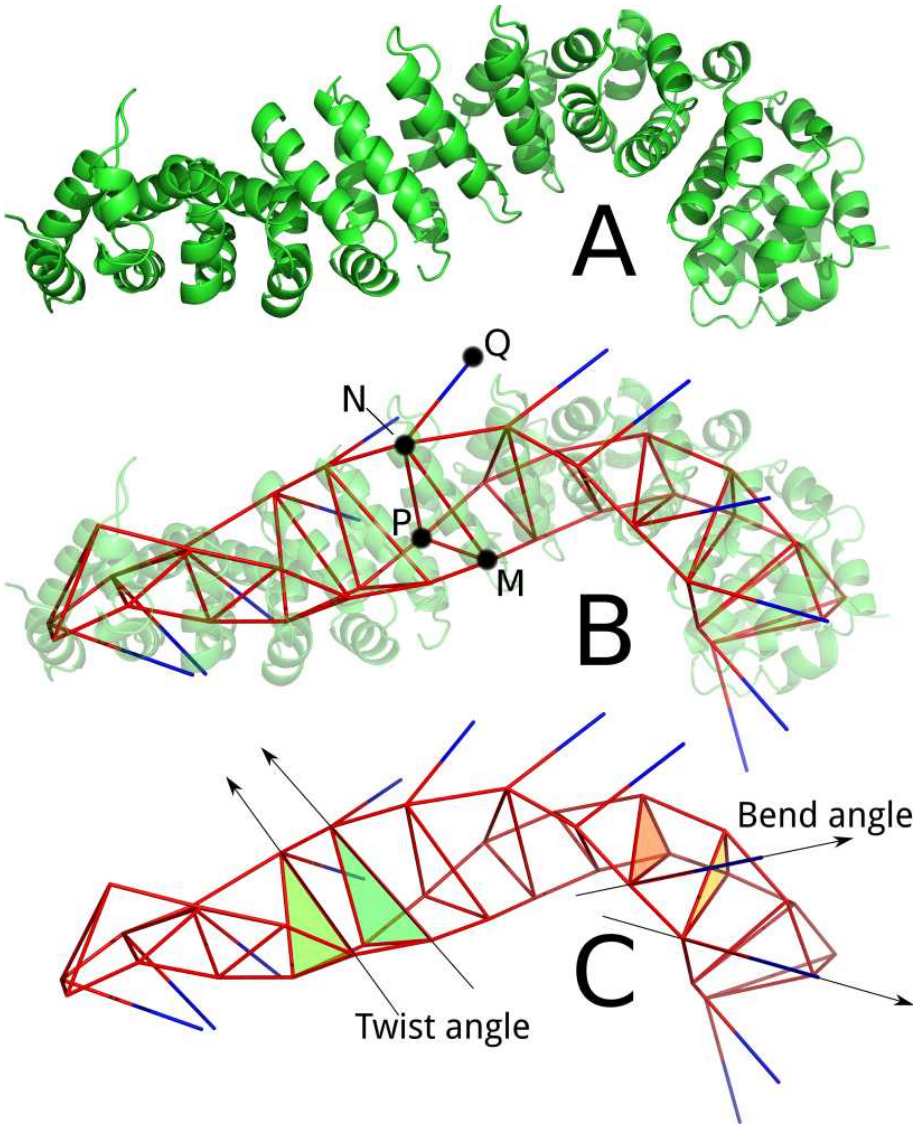


Figure 18: **Geometrical analysis.** Figure A shows an armadillo repeat protein in cartoon representation. Figure B shows how the internal repeats are considered for geometrical analysis. The segment $\overline{\text{NM}}$ is the axis of the helix 3 of a repeat. Helix 3 is the longest and contains the asparagine responsible for binding the backbone of the interacting peptide. The point P is the center of gravity of the helices 1 and 2. The triangle NMP approximates a plane on which the repeat lies, and the segment $\overline{\text{NQ}}$ is normal to it. Figure C shows how geometrical parameters are computed. The interrepeat bend angle is the angle between the normals ($\overline{\text{NQ}}$) of repeat dimers; the interrepeat twist angle is the angle between the axes of the helix 3 ($\overline{\text{NM}}$) of repeat dimers.

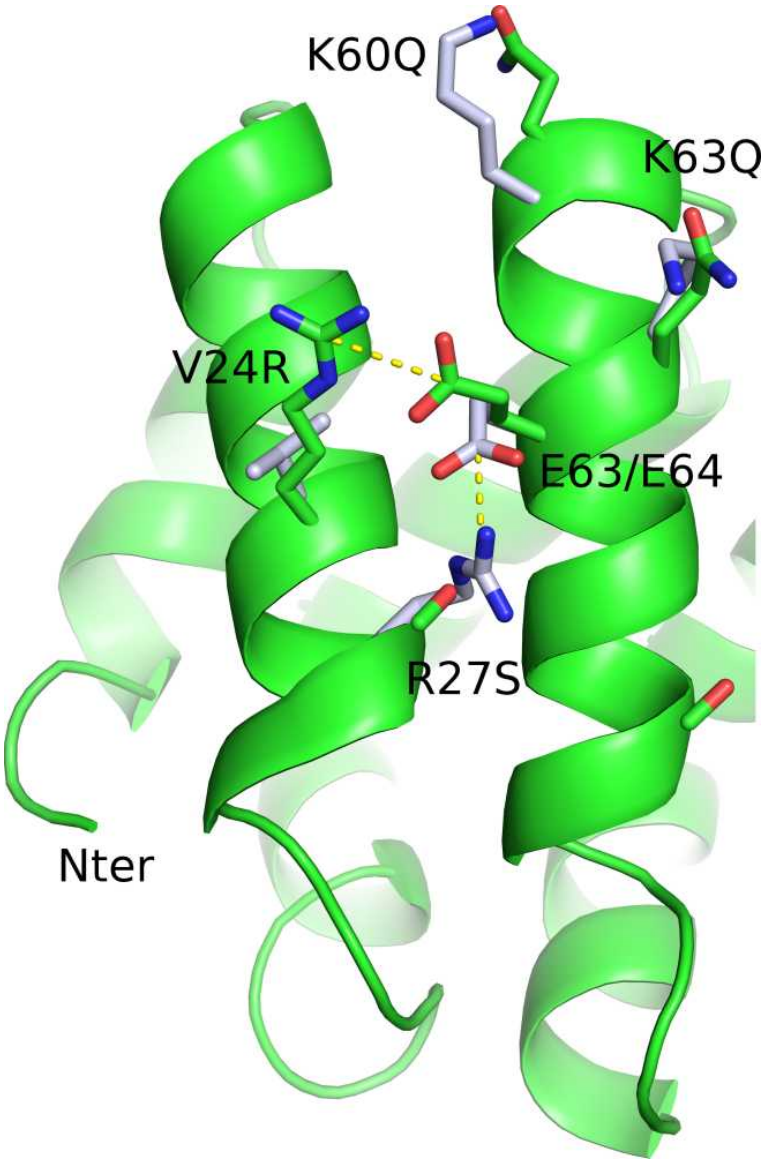


Figure 19: **The interrepeat salt bridge and mutations at the Ncap.** The protein shown in green cartoon is the conformation of Δ5 used for starting explicit solvent simulations. Residues colored in green are from Δ5, while residues colored in white are from KK, after structural superposition of Ncap-R1. KK conformation is omitted for clarity reasons. The figure shows the mutations described in the text and listed in table 2. K60 and K63 are the lysines of the first repeat mutated to glutammines in the QQ and Δ5 models. The salt bridge between R27 and E64 (in KK and QQ) and R24 and E63 (in Δ5) is shown as dashed lines.

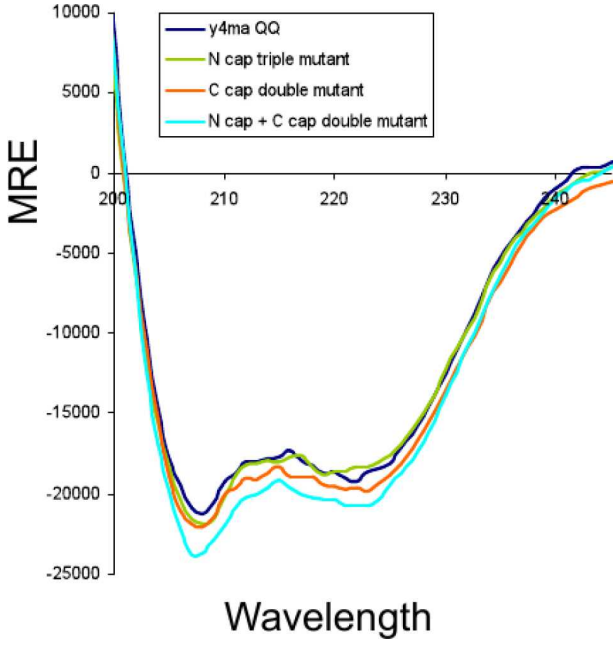


Figure 20: **Armadillo repeat proteins have a helical secondary structure**. CD spectra of YM_4A and several mutants show the characteristic signal of helices.

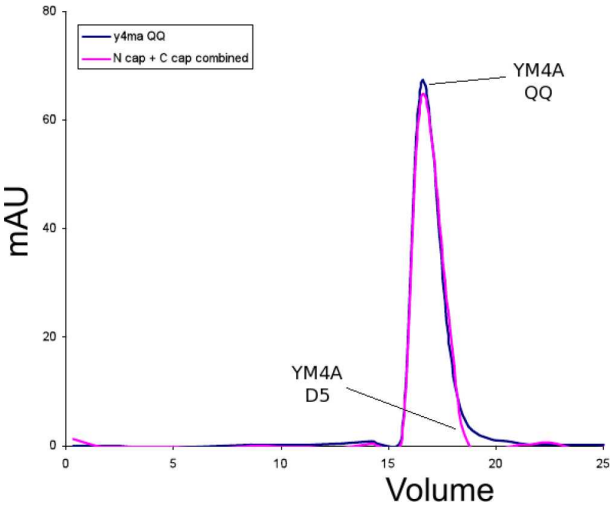


Figure 21: **YM_4A QQ and $\Delta 5$ are monomeric species**. Size-exclusion chromatography plots show that YM_4A QQ and $\Delta 5$ are eluited at two very close volumes and they are monomeric species. Other mutants are monomeric species, too (data not shown). MALS analysis confirmed the SEC results on all constructs.

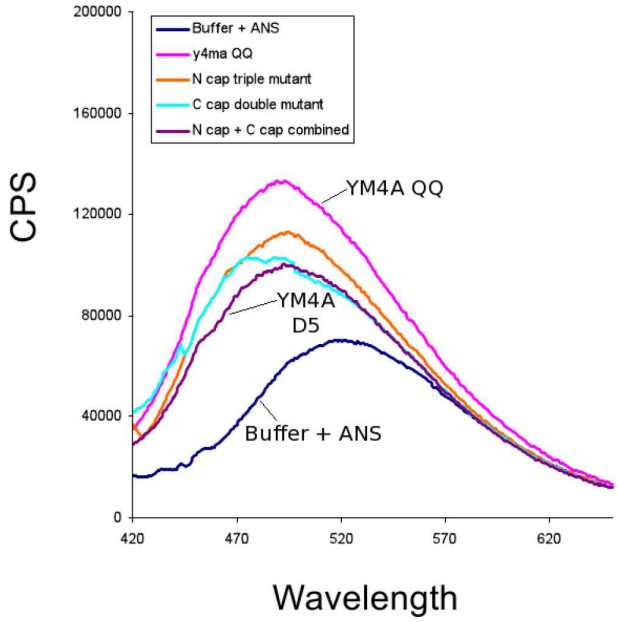


Figure 22: **Reduced hydrophobic surface of Δ5**. The fluorescent dye ANS (1-Anilino-8-naphthalene sulfonate) binds to solvent exposed hydrophobic patches of proteins. The difference between the curve of QQ and Δ5 shows that the mutations reduce the hydrophobic solvent exposed surface.

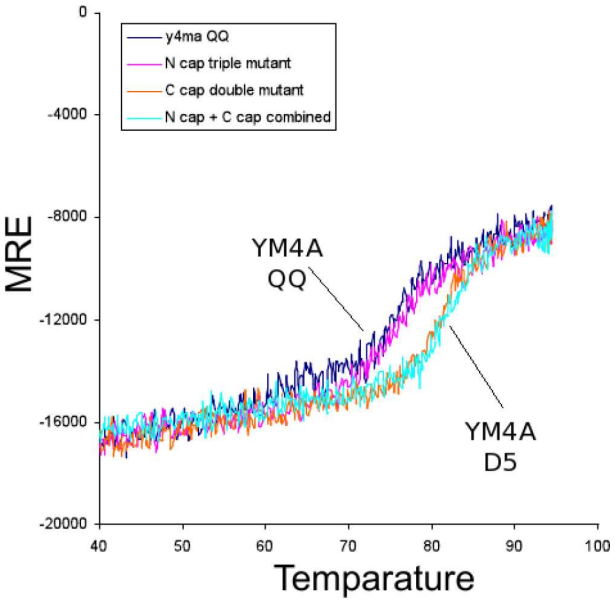


Figure 23: **Improved thermal stability of $\Delta 5$**. The temperature of melting of YM_4A QQ deduced from the thermal denaturation curves is 77 Celsius degrees. The temperature of melting of the N-cap mutant, the C-cap mutant, and of $\Delta 5$ is 78.5, 85.0, and 86.3 Celsius degrees, respectively.

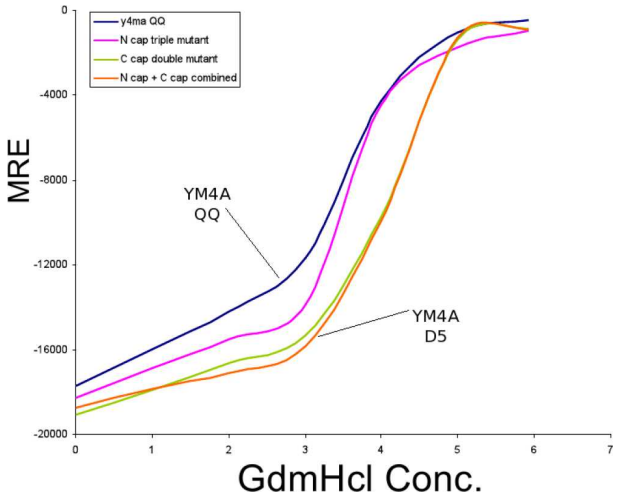


Figure 24: **Improved chemical stability of Δ5**. The Δ5 and the N-cap mutants denaturate at higher concentrations of guanidinium chloride with respect to YM_4A QQ.

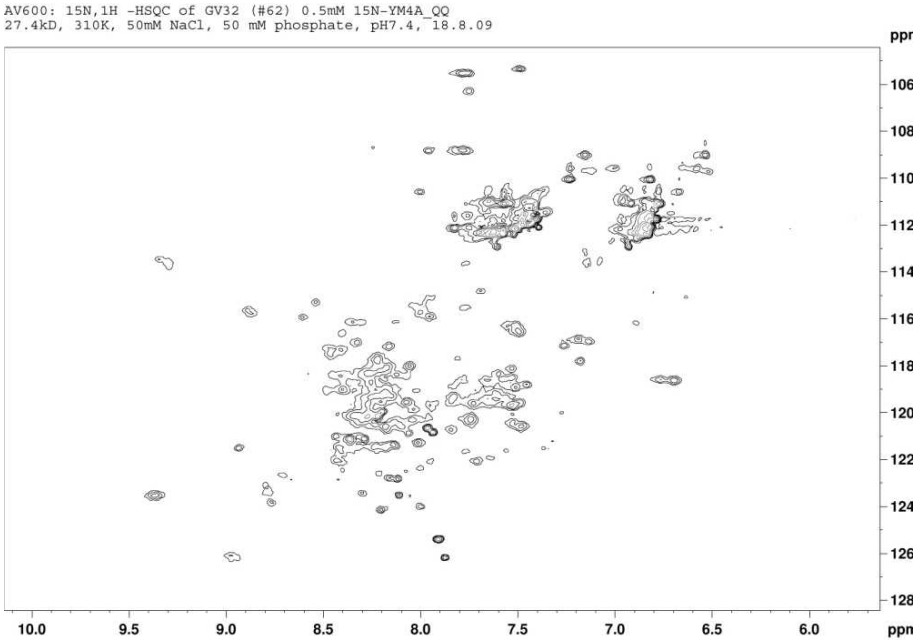


Figure 25: **2D-NMR spectrum of YM_4A QQ**. The dispersion and intensity of the peaks is lower than in the spectrum of YM_4A $\Delta 5$ (see figure 26).

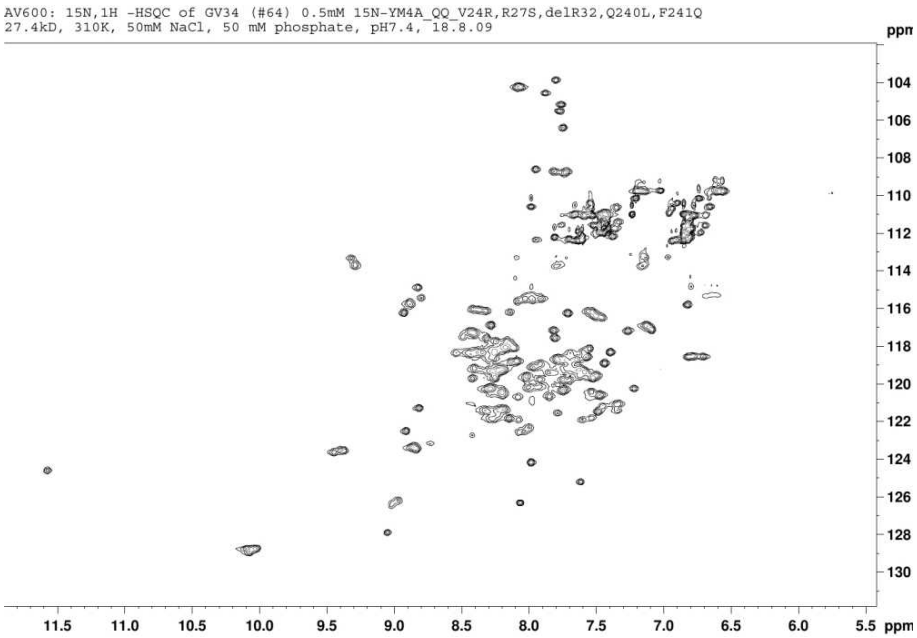


Figure 26: **2D-NMR spectrum of YM_4A $\Delta 5$.**

6 Tables

original PDB code	number of internal repeats		simulation time
1EE4	8	yeast importin α	48ns
1Q1T	8	mouse importin α	57ns
2BCT	10	murine β-catenin	35ns

Table 1: **Simulations run on starting models.**

Original sequence	Mutations	Label
YM_4A	–	KK
YM_4A	K60 K102Q K144Q K186Q K63Q K105Q K147Q K189Q	QQ
QQ	Q240L	Q240L
QQ	R27S V24R ΔR32	Ncap
QQ	R27S V24R ΔR32 Q240L F241Q	Δ5

Table 2: **List of run simulations and studied mutations.**

YM_3A	QQ		
	QQ	V24R R27S ΔR32	
	QQ	Q240L F241Q	
	QQ	V24R R27S ΔR32	Q240L F241Q
YM_4A	QQ		
	QQ	V24R R27S ΔR32	
	QQ	Q240L F241Q	
	QQ	V24R R27S ΔR32	Q240L F241Q

Table 3: **Mutants analyzed by NMR experiments.**

		KK			QQ			Δ5		
R1-R4	1EE4	54	0	0	34	0	15	49	0	0
	1Q1T	76	0	0	20	0	0	48	0	0
	2BCT	0	0	0	0	0	0	0	0	0
R1-R3 R2-R4	1EE4	83	0	0	92	0	16	76	0	0
	1Q1T	83	0	0	73	0	0	70	0	0
	2BCT	8	2	2	17	0	0	27	0	0

Table 4: **Superposition of models to the three X-ray structures used for homology models.** For the KK, QQ, and Δ5 model, the three colums shows data from three explicit water simulations. The numbers indicate the percentage of the simulation time in which the RMSD to a particular X-ray structure (1EE4, 1Q1T, and 2BCT) is lower than 2 Å. These three X-ray structures were used for generating the initial three homology models. The label R1-R4 refers to the superposition of all four internal repeats to the X-ray structure, while the labels R1-R3 and R2-R4 refer to the superposition of the first three or the last three repeats, respectively. The superposition is carried out on the whole x-ray structure. For the three internal repeats superposition, only the best value is shown. If only three repeats are considered, the criteria for similarity is less strict.

References

[1] I. Andricioaei and M. Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, 115:6289, 2001.

[2] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), 1983.

[3] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98:10089, 1993.

[4] M. Gerstein and M. Levitt. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Science*, 7(2):445–456, 1998.

[5] U. Haberthur and A. Caflisch. FACTS: Fast analytical continuum treatment of solvation. *Journal of computational chemistry*, 29(5):701, 2008.

[6] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31:1695, 1985.

[7] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983.

[8] A.D. Mackerell, M. Feig, and C.L. Brooks. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry*, 25(11):1400–1415, 2004.

[9] AD MacKerell Jr, D. Bashford, M. Bellott, RL Dunbrack Jr, JD Evanseck, MJ Field, S. Fischer, J. Gao, H. Guo, S. Ha, et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998.

[10] L. Nilsson. Efficient table lookup without inverse square roots for calculation of pair wise atomic interactions in classical simulations. *Journal of Computational Chemistry*, 30(9), 2009.

[11] S. Nose. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81:511, 1984.

[12] F. Parmeggiani, R. Pellarin, A.P. Larsen, G. Varadamsetty, M.T. Stumpp, O. Zerbe, A. Caflisch, and A. Plückthun. Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. *Journal of Molecular Biology*, 2007.

[13] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch. Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics*, 23(19):2625, 2007.

7 Supplementary material

7.1 Complete RMSF plots of implicit and explicit solvent simulations

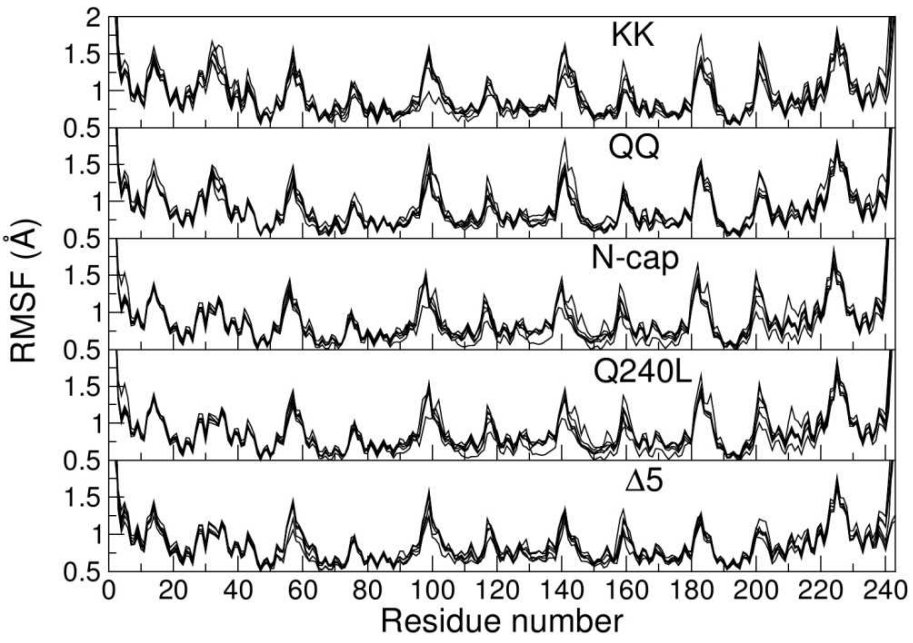


Figure 27: **RMSF plot superposition of all fifth generation implicit water simulations.**

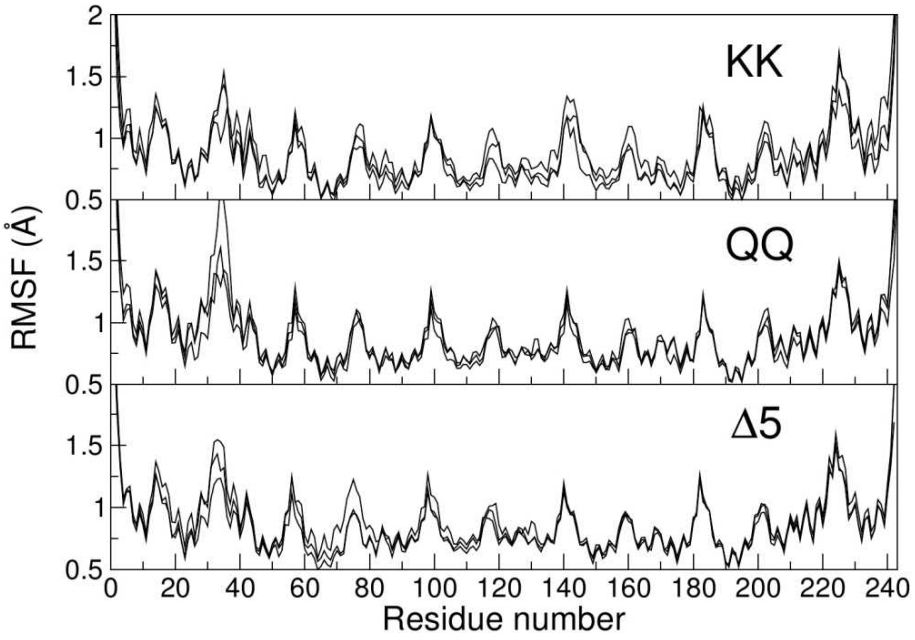


Figure 28: **RMSF plot superposition of all explicit water simulations.**

7.2 Bend and twist angles timeseries

Bend and twist angles were calculated as described in Materials and Methods. The "Asn Cβ distance" is the distance between two asparagine Cβ atoms of consecutive repeats. The asparagines occupy the same position repeat wise and are responsible for interacting with the backbone of the bound peptide in the crystal structures. The optimal values observed in the crystal structures for binding the peptide are the two horizonal red lines. The red lines in the bend and twist angles plots show the range in which the most common values for the respective measured quantities lie in the three crystal structures used for generating the homology models.

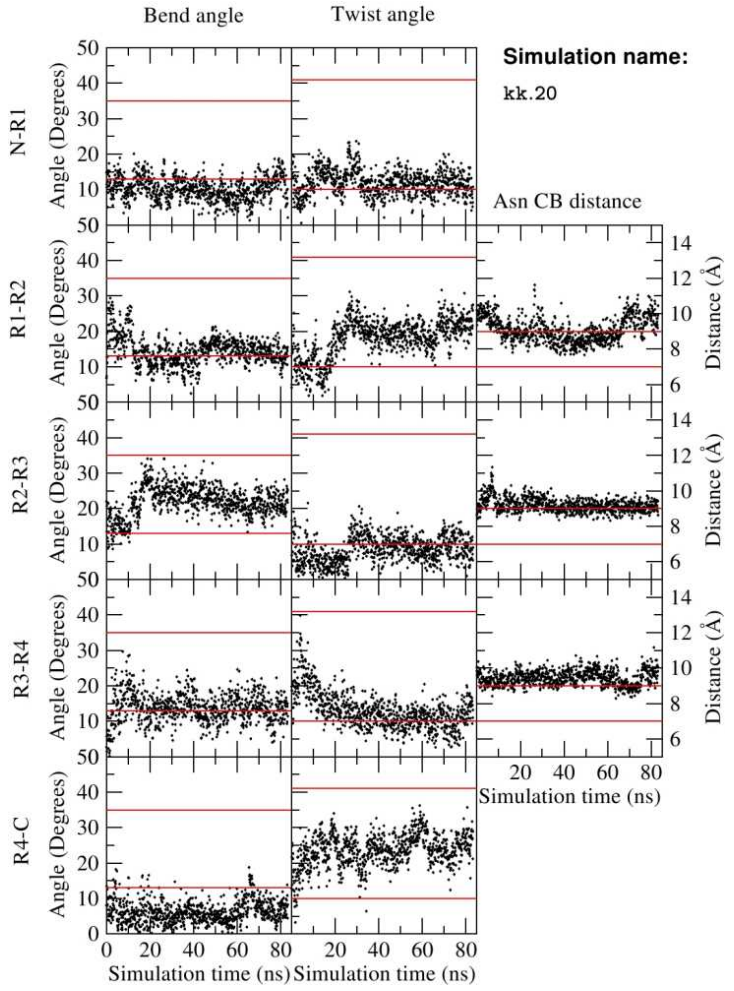


Figure 29: **Bend and twist angles timeseries of the first KK model simulation**

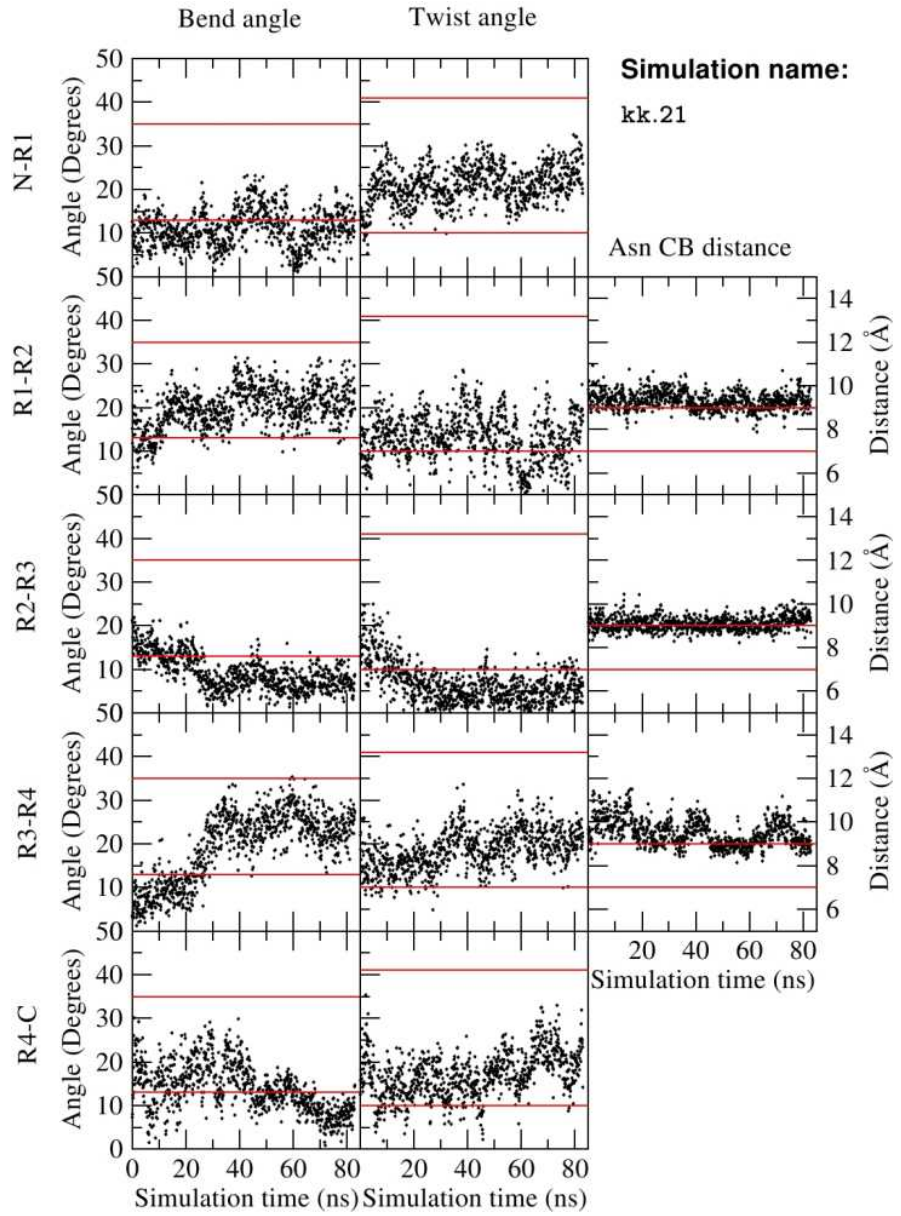


Figure 30: **Bend and twist angles timeseries of the second KK model simulation**

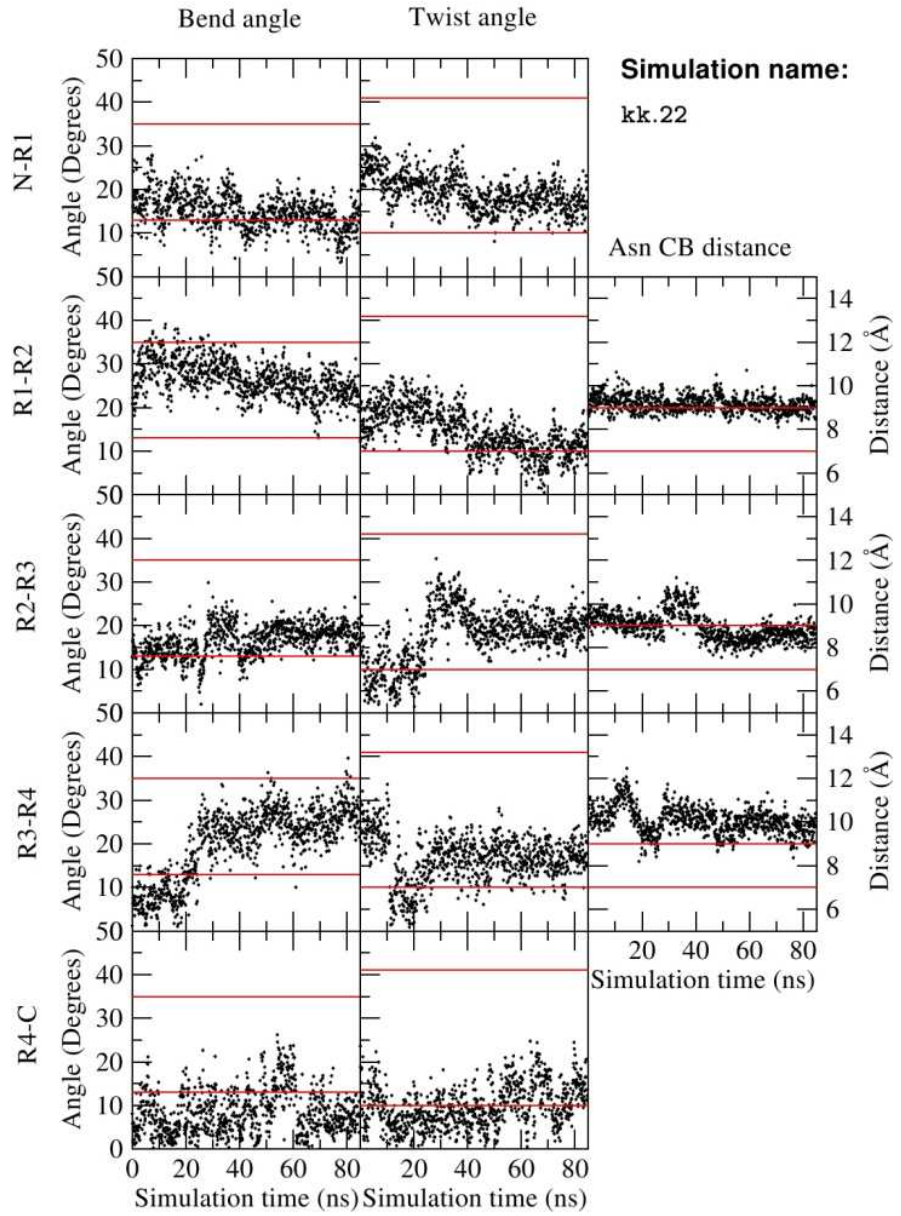


Figure 31: **Bend and twist angles timeseries of the third KK model simulation**

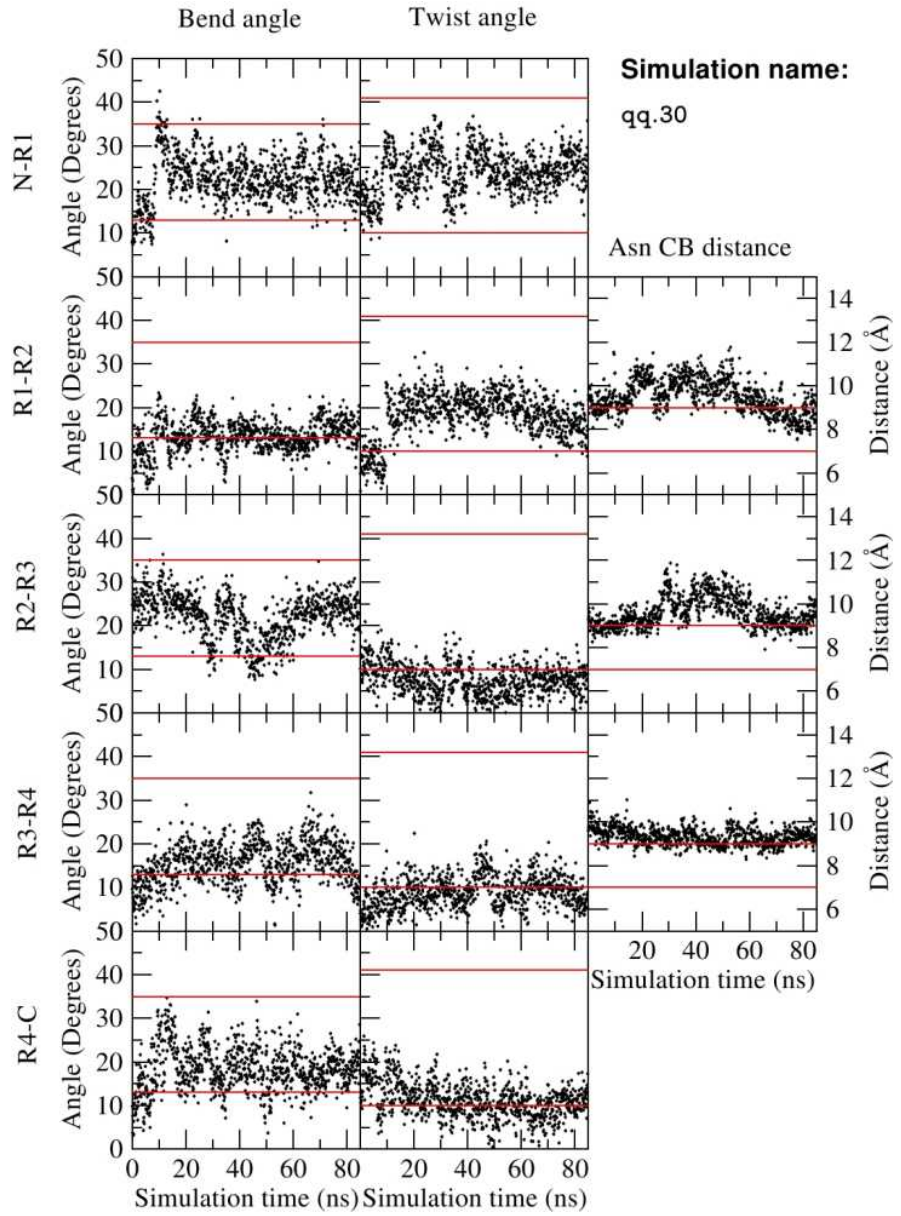


Figure 32: **Bend and twist angles timeseries of the first QQ model simulation**

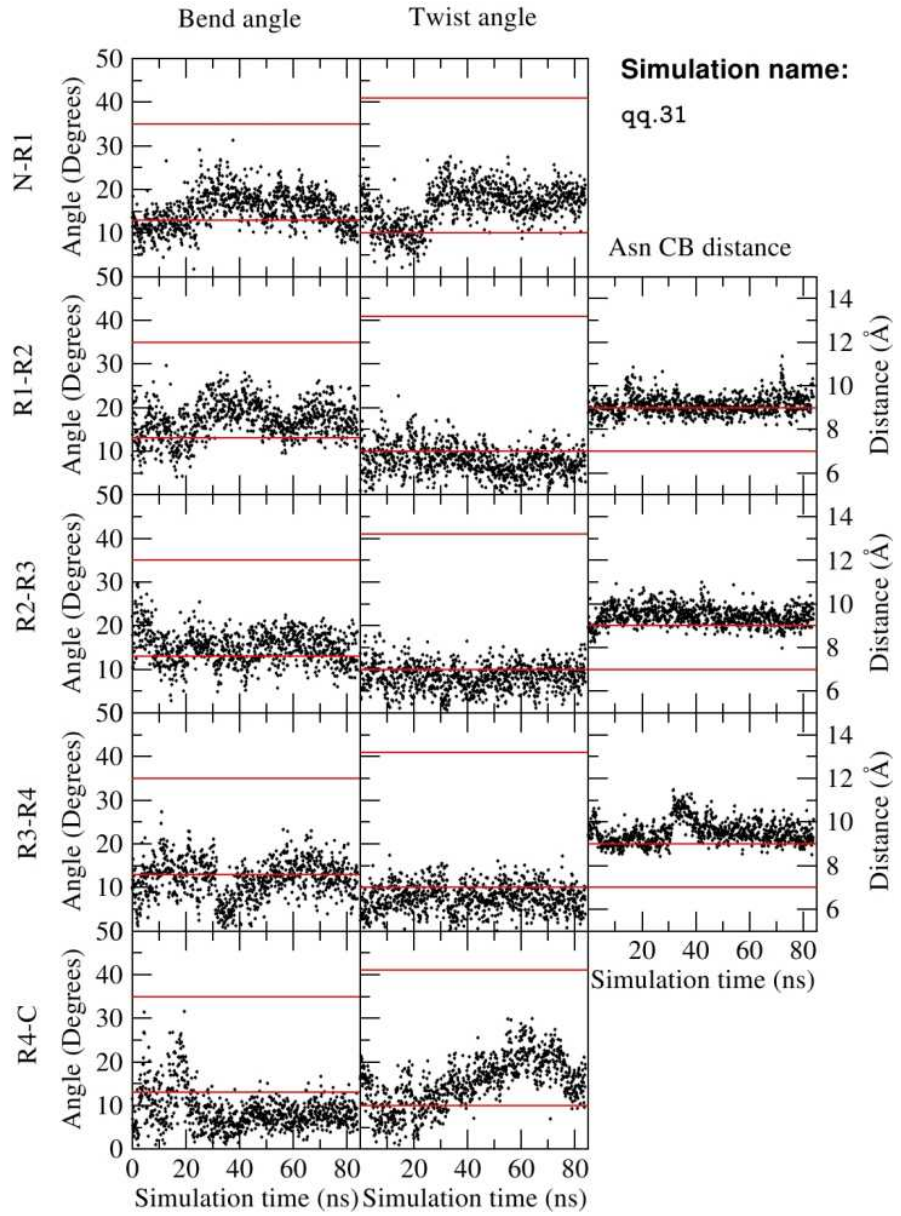


Figure 33: **Bend and twist angles timeseries of the second QQ model simulation**

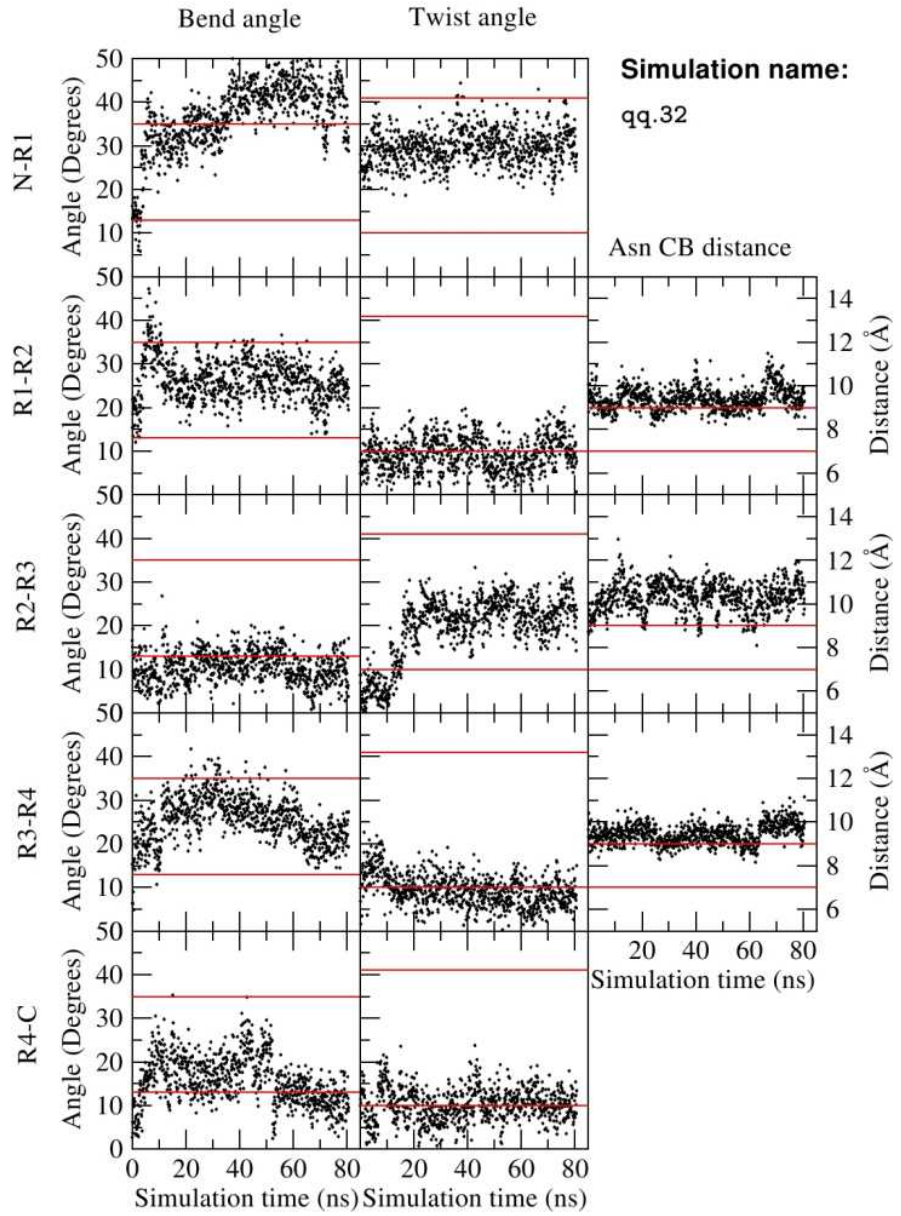


Figure 34: **Bend and twist angles timeseries of the third QQ model simulation**

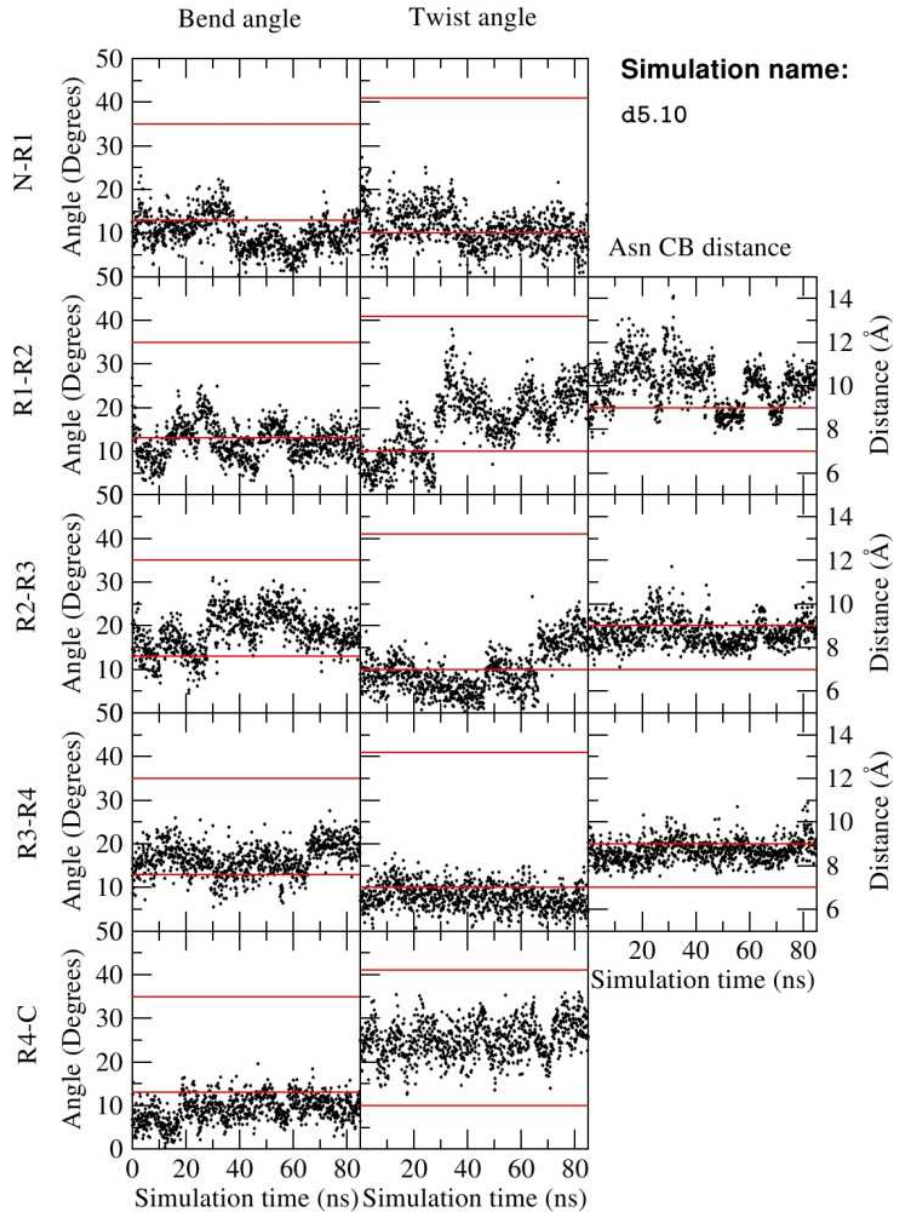


Figure 35: **Bend and twist angles timeseries of the first Δ5 model simulation**

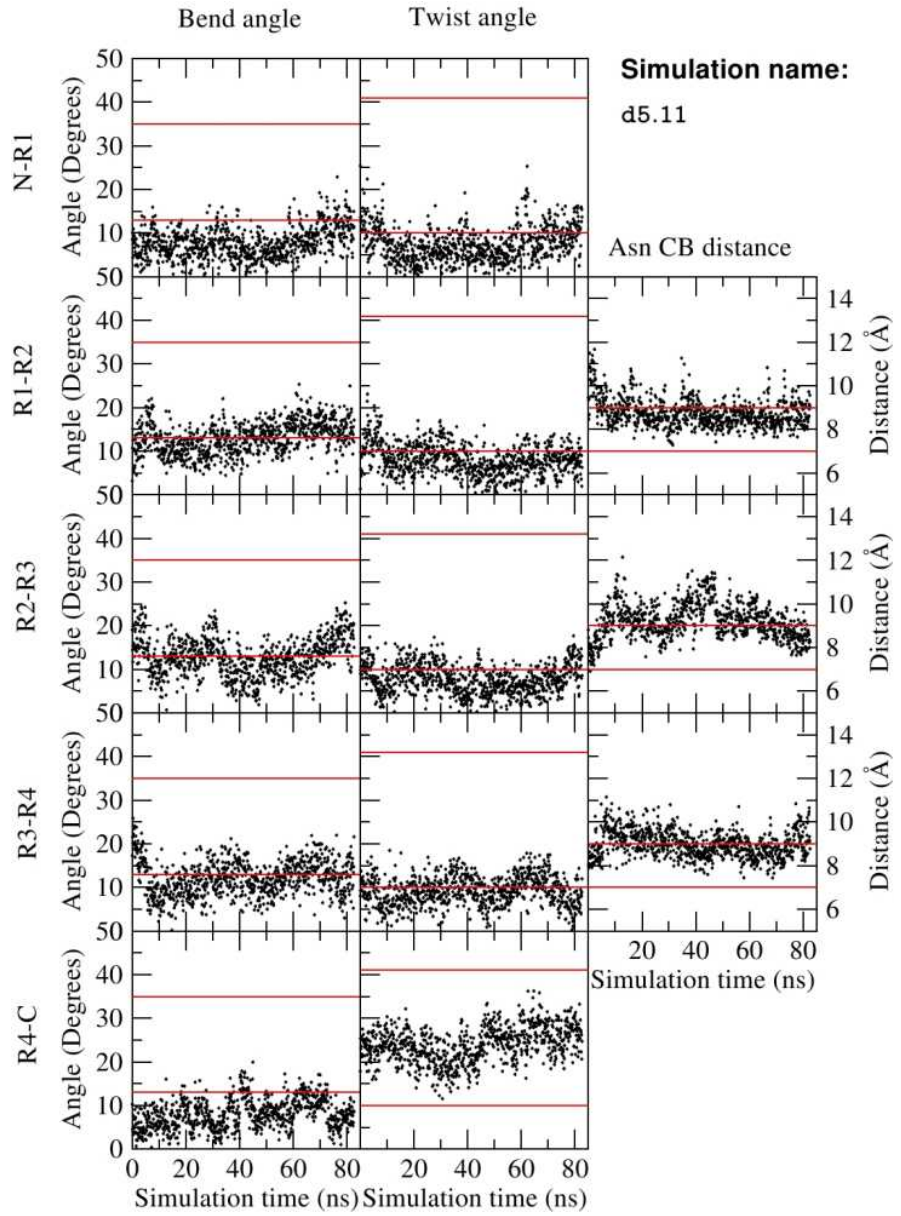


Figure 36: **Bend and twist angles timeseries of the second Δ5 model simulation**

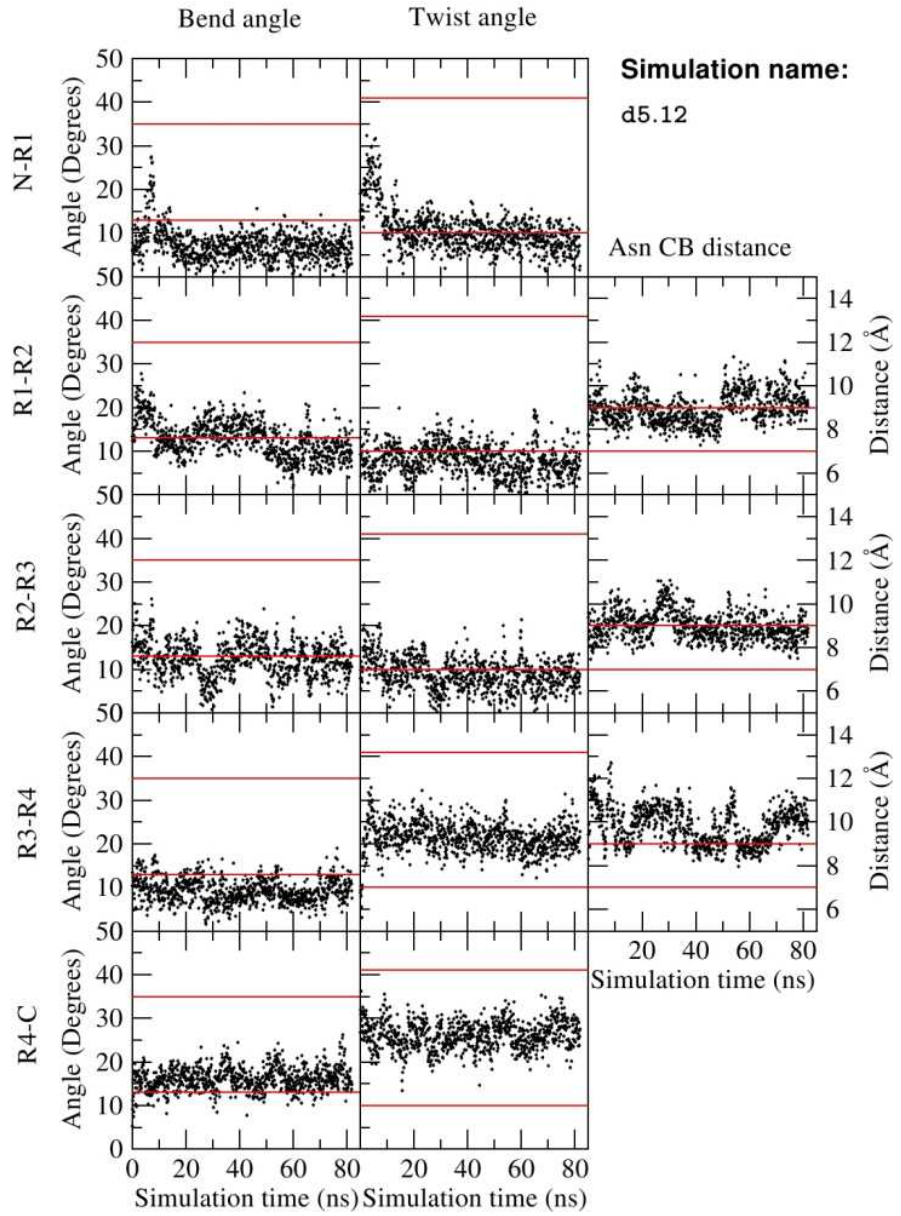


Figure 37: **Bend and twist angles timeseries of the third Δ5 model simulation**

7.3 RMSD to X-ray structures timeseries

The black timeseries is the RMSD and the red timeseries is the percentage of kept atom pairs from the initial alignment according to the *fit structal* algorithm of Witnotp (see Material and Methods). To be plotted in the same range of the RMSD, the percentage of kept atom pairs has been normalized to 8, therefore 100% kept atom pairs is equal to 8.

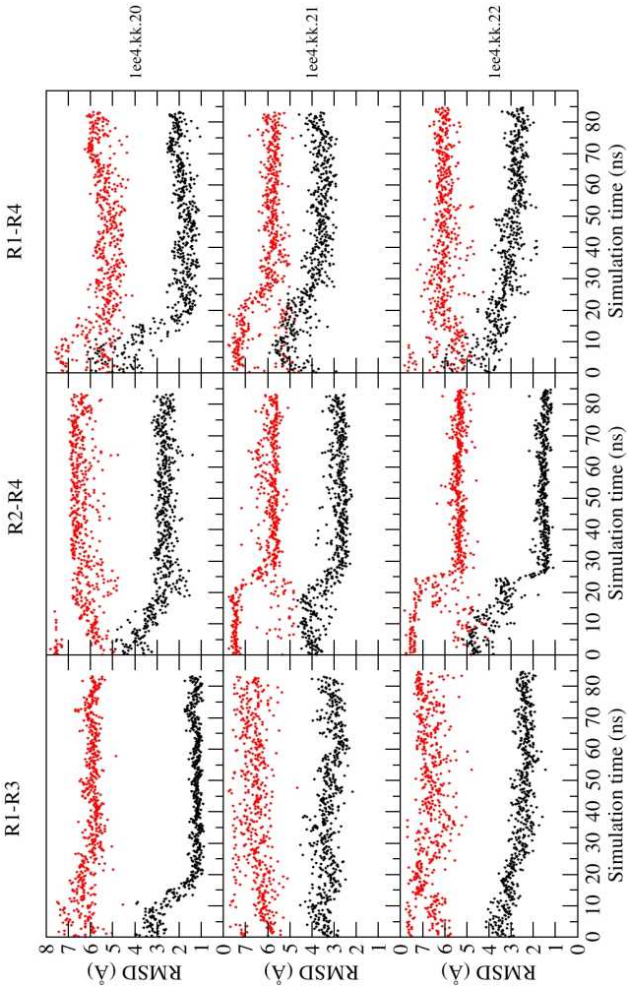


Figure 38: **RMSD superposition of KK model on the X-ray structure 1EE4**.

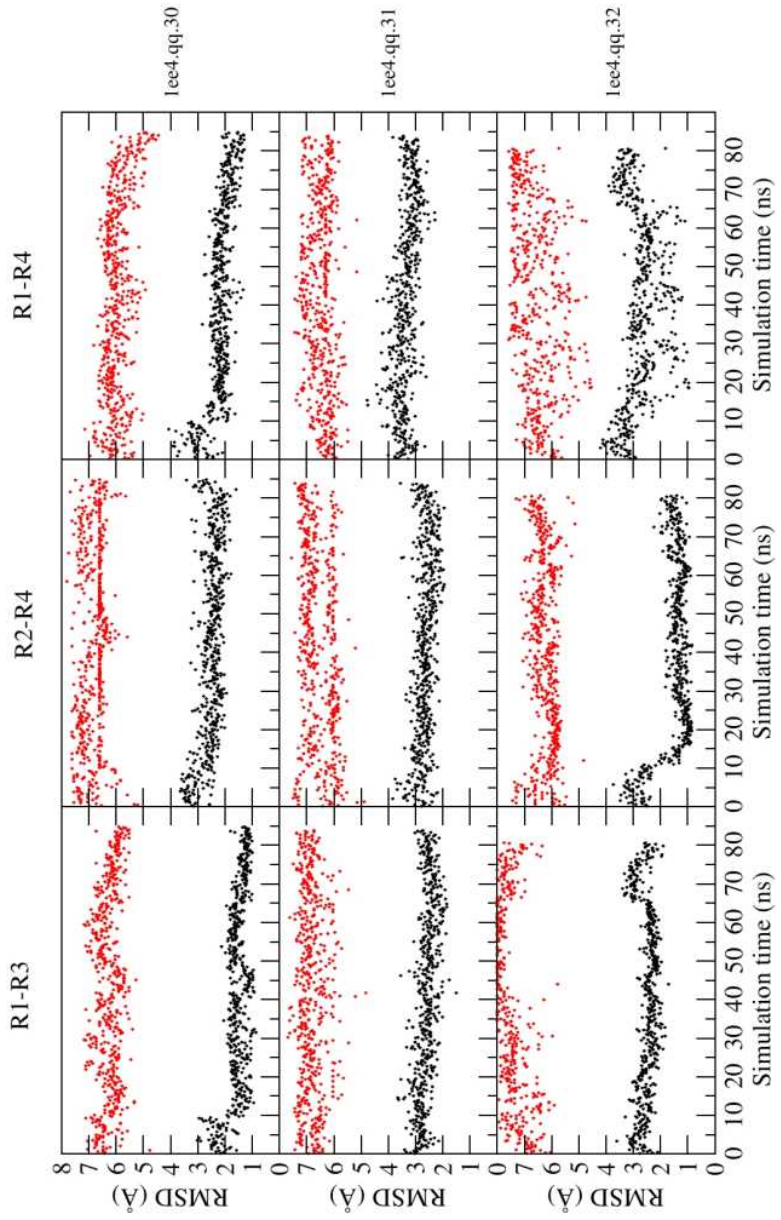


Figure 39: **RMSD superposition of QQ model on the X-ray structure 1EE4**

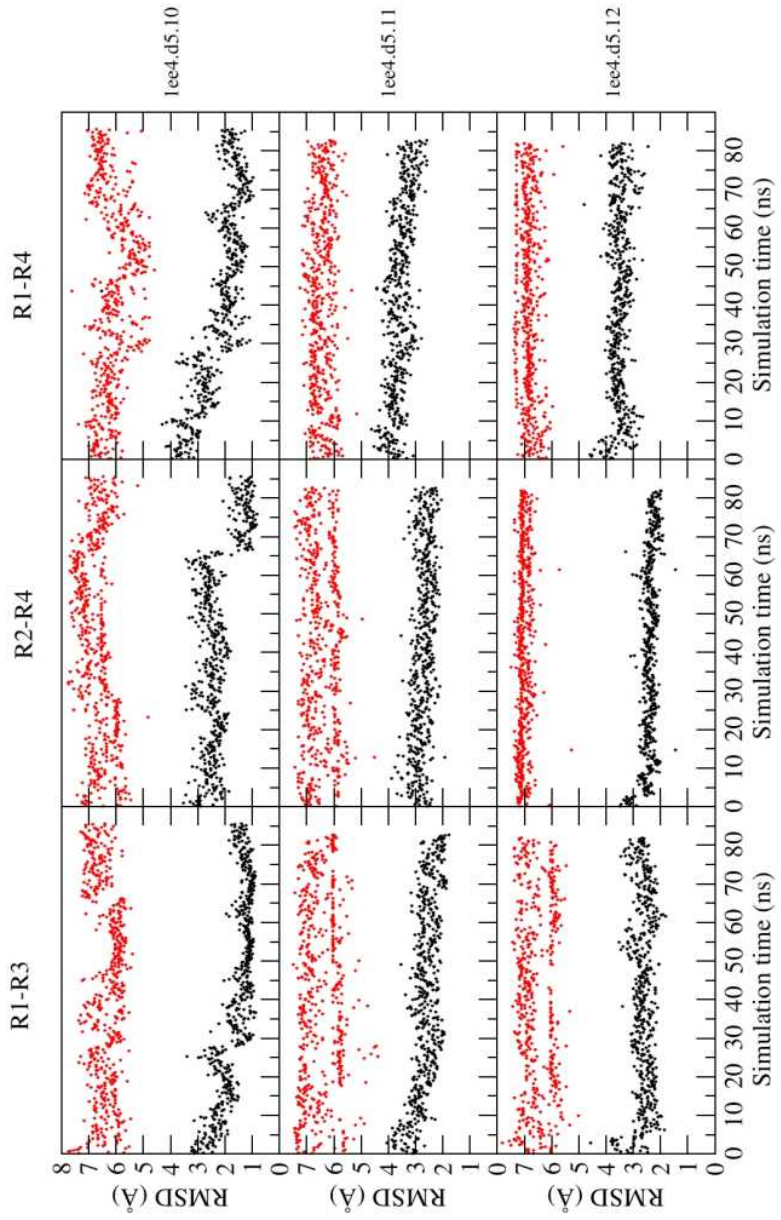


Figure 40: **RMSD superposition of Δ5 model on the X-ray structure 1EE4**

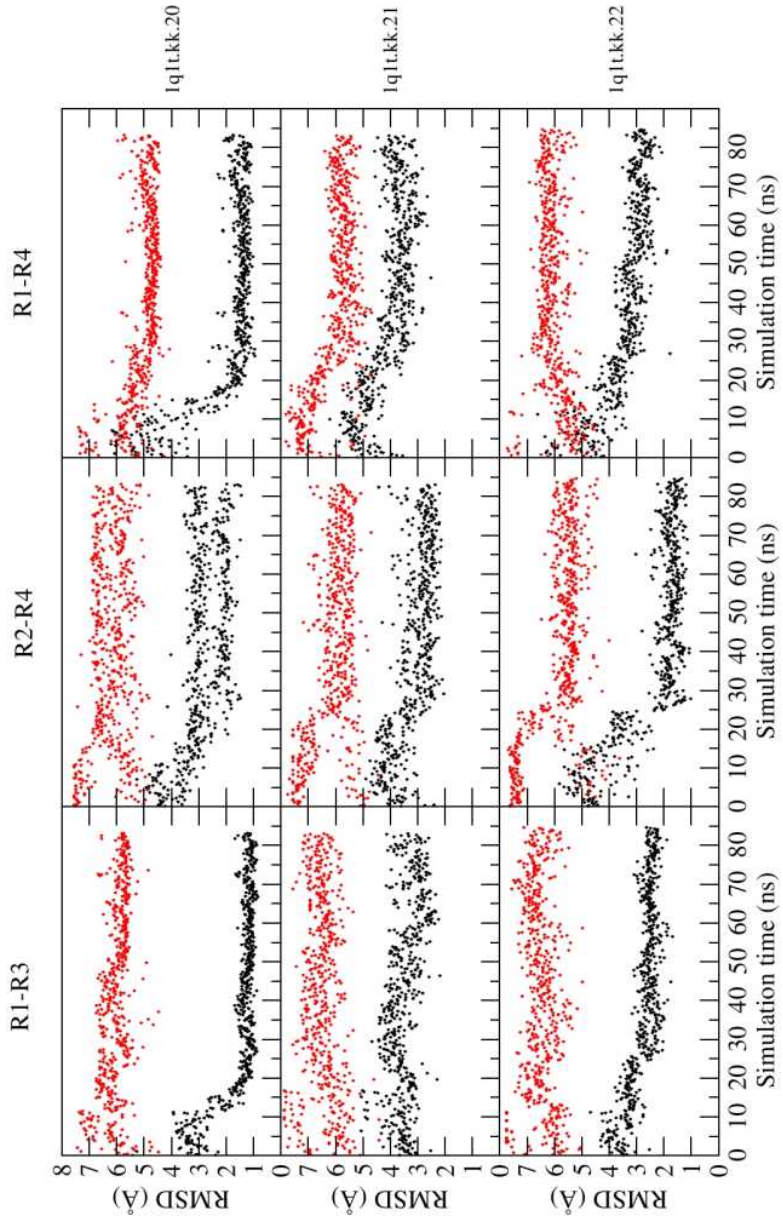


Figure 41: **RMSD superposition of KK model on the X-ray structure 1Q1T**.

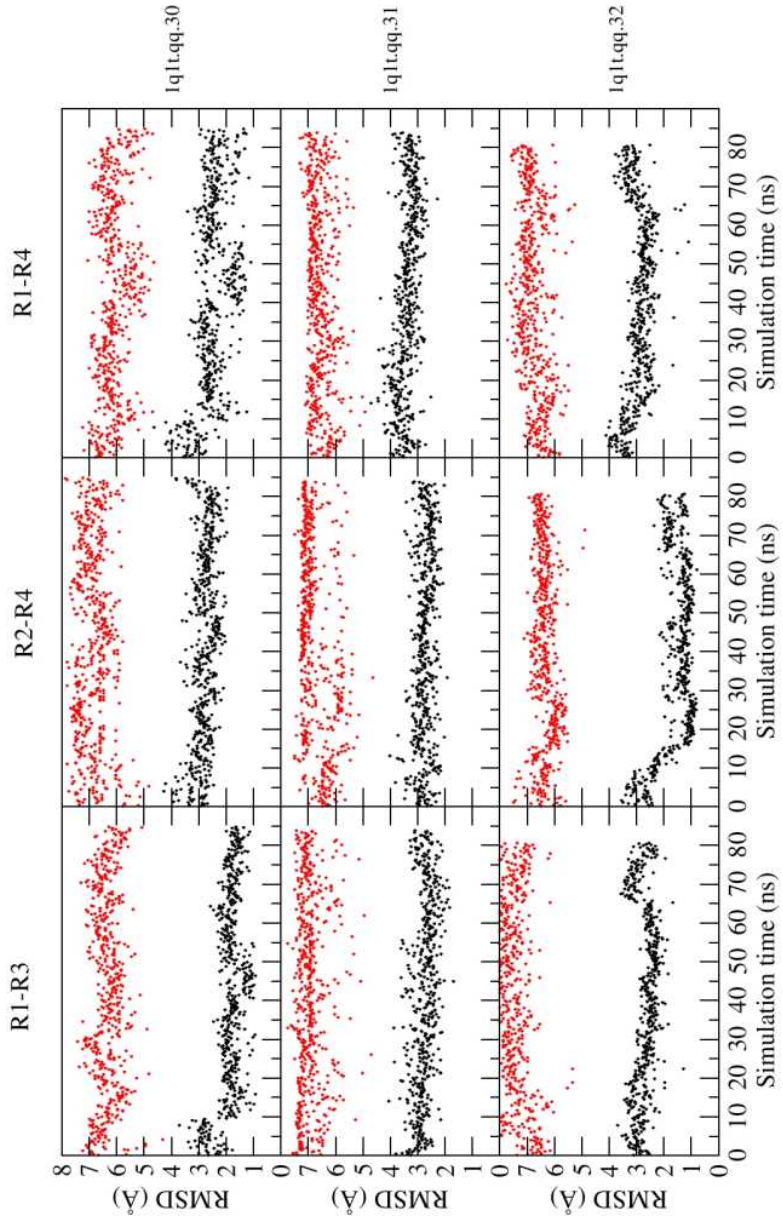


Figure 42: **RMSD superposition of QQ model on the X-ray structure 1Q1T**

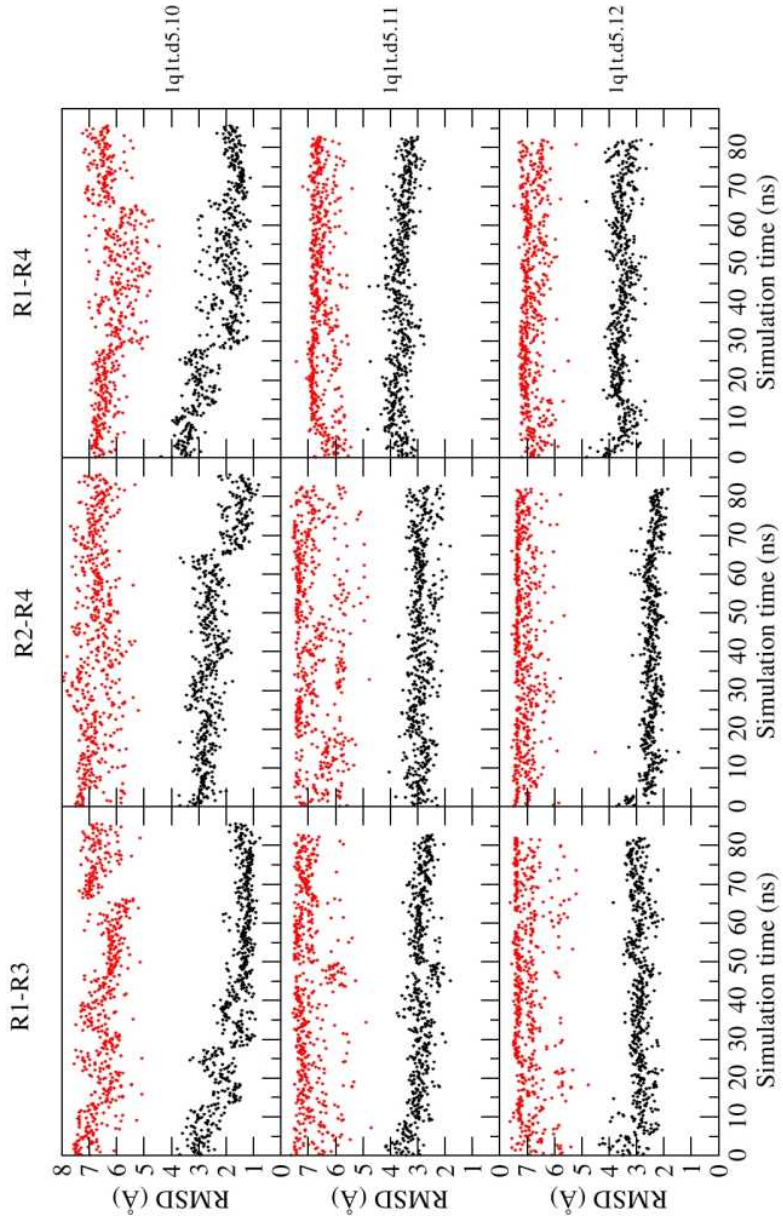


Figure 43: **RMSD superposition of Δ5 model on the X-ray structure 1Q1T**

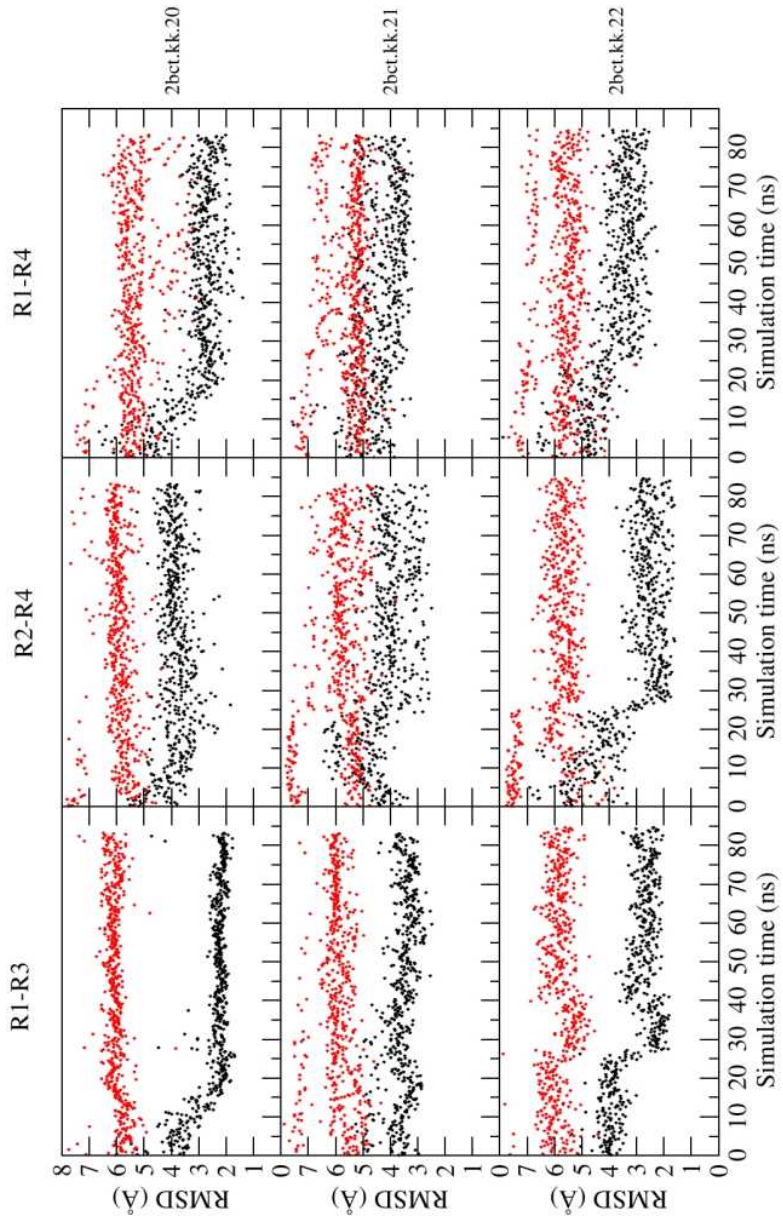


Figure 44: **RMSD superposition of KK model on the X-ray structure 2BCT**.

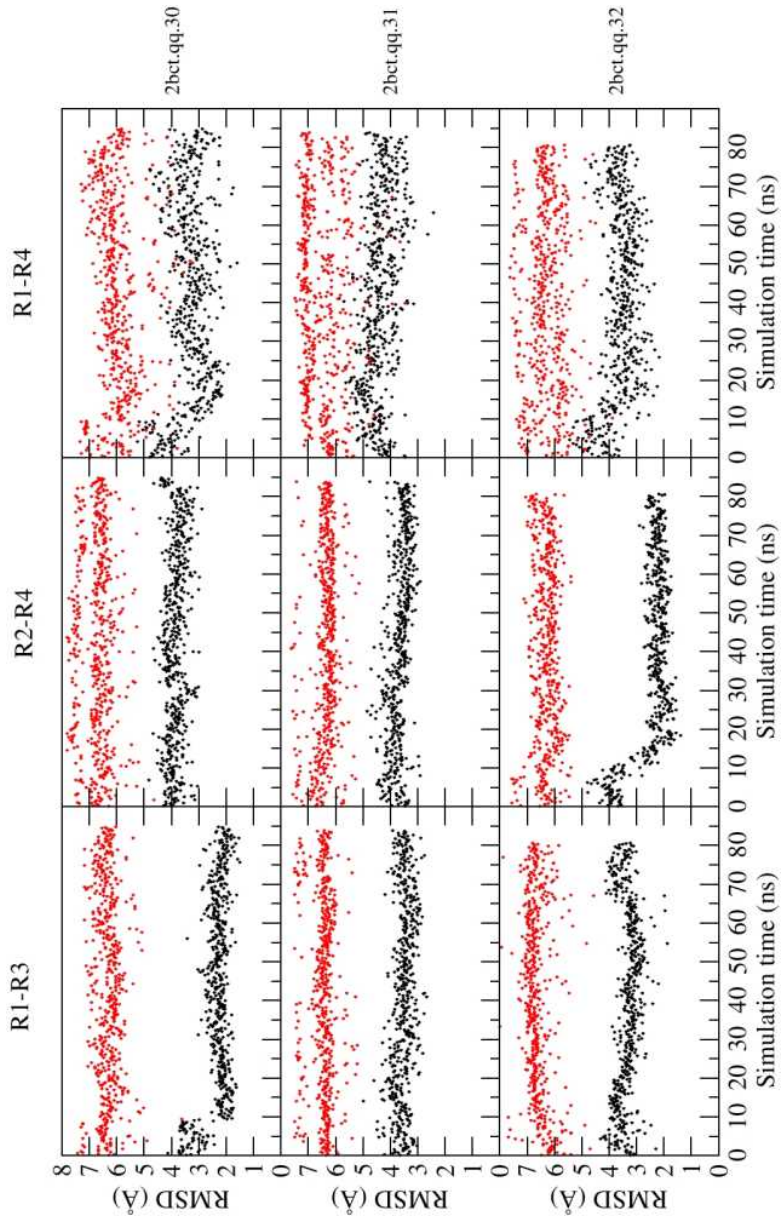


Figure 45: **RMSD superposition of QQ model on the X-ray structure 2BCT**

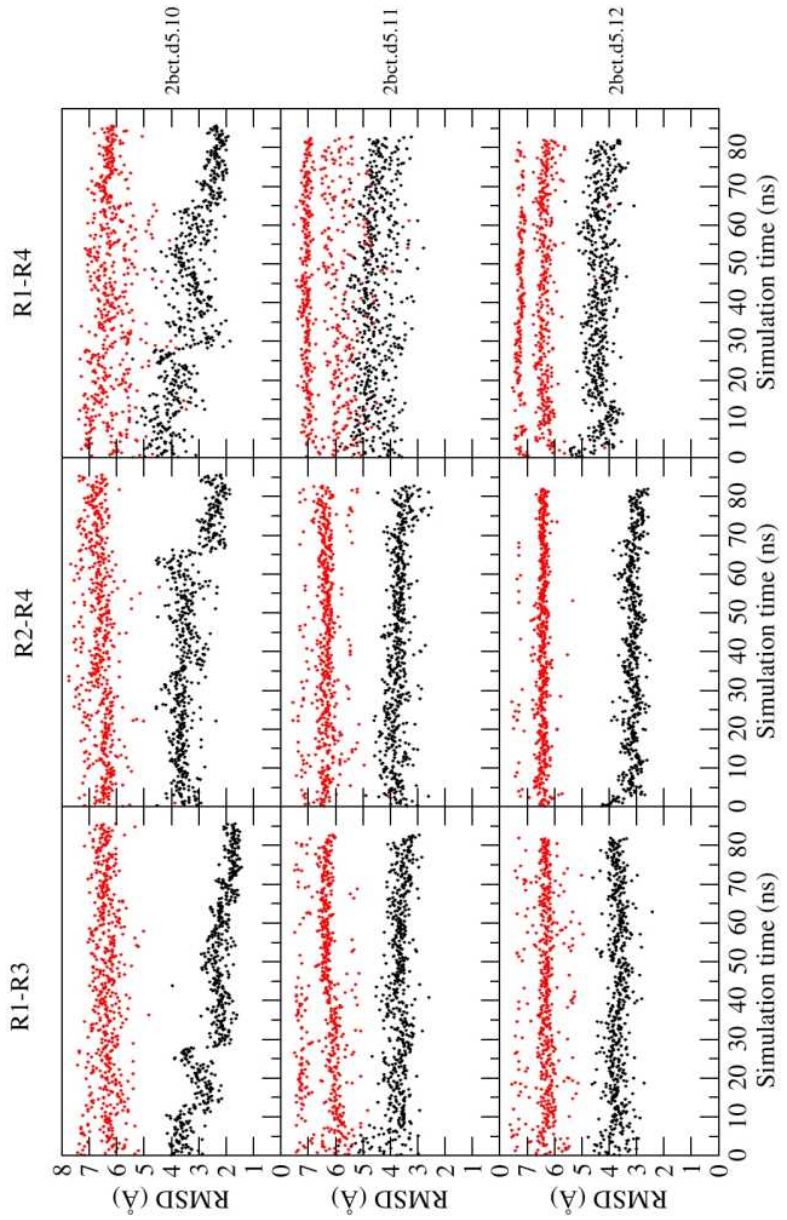


Figure 46: **RMSD superposition of Δ5 model on the X-ray structure 2BCT**

Chapter 7

Conclusions

Structure-based methods were successfully applied to fields such as virtual screening (cathepsin B and cathepsin K projects), protein design (Armadillo repeat proteins project), and protein function studies (Chromodomain and Cryptochrome projects).

Considering all the different projects, I can surely say that the *leit motiv* of my doctoral thesis has been protein flexibility and dynamics.

The cathepsin B project showed that protein dynamics, such as the opening and closing of the occluding loop, which were hidden in the static representation of a crystal structure, could be exploited for modifying the activity of the enzyme. Docking of a library of compounds was employed for finding molecules which could interact with the loop. A compound found by docking was experimentally confirmed to be an inhibitor of cathepsin B.

Protein flexibility and dynamics played a major role in the cathepsin K project, where a predicted putative allosteric binding pocket was independently confirmed by two methods, normal mode analysis and principal component analysis, which rely on the flexible nature of proteins. A virtual screening campaign by docking a library of small molecules to this pocket was started and a series of activators of cathepsin K activators were found. Further experiments, such as the determination of the X-ray structure of the complex, are needed for the confirmation of the binding pocket.

Understanding the protein dynamics was the basis for the optimization of the artificial armadillo repeat protein YM_4A. The dispersion of the NMR signals, which depends on the flexibility of the protein, was markedly improved by introducing a few mutations that reduced the flexibility of the protein in molecular dynamics simulations.

The alteration of protein flexibility by a single point mutation, found from the analysis of molecular dynamics trajectories, contributed to the explanation of the difference of activity of a Cryptochrome of Arabidopsis.

These results altogether show an application of computer-based methods, not only in drug-discovery related fields, but also in basic research.

Bibliography

[1] J. Aqvist and J. Marelius. The linear interaction energy method for predicting ligand binding free energies. *Combinatorial Chemistry & High Throughput Screening*, 4(8):613–626, 2001.

[2] I. Bahar, T.R. Lezon, A. Bakan, and I.H. Shrivastava. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chemical reviews*, pages 1–432, 2009.

[3] C. Bissantz, G. Folkers, and D. Rognan. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem*, 43(25):4759–4767, 2000.

[4] C. Bissantz, B. Kuhn, and M. Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry*, 2010, article ASAP.

[5] J. Bostrom, A. Hogner, and S. Schmitt. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem*, 49(23):6716–6725, 2006.

[6] BR Brooks, CL Brooks III, AD Mackerell Jr, L. Nilsson, RJ Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.

[7] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), 1983.

[8] N. Budin, N. Majeux, and A. Caflisch. Fragment-based flexible ligand docking by evolutionary optimization. *Biological Chemistry*, 382(9):1365–1372, 2001.

[9] G. Campiani, A.P. Kozikowski, S. Wang, L. Ming, V. Nacci, A. Saxena, and B.P. Doctor. Synthesis and anticholinesterase activity of huperzine

A analogues containing phenol and catechol replacements for the pyridone ring. *Bioorganic & medicinal chemistry letters*, 8(11):1413–1418, 1998.

[10] R.A.E. Carr, M. Congreve, C.W. Murray, and D.C. Rees. Fragment-based lead discovery: leads by design. *Drug discovery today*, 10(14):987–992, 2005.

[11] C.N. Cavasotto and R.A. Abagyan. Protein flexibility in ligand docking and virtual screening to protein kinases. *Journal of molecular biology*, 337(1):209–225, 2004.

[12] C.N. Cavasotto, J.A. Kovacs, and R.A. Abagyan. Representing receptor flexibility in ligand docking through relevant normal modes. *J. Am. Chem. Soc*, 127(26):9632–9640, 2005.

[13] M. Cecchini, A. Houdusse, and M. Karplus. Allosteric communication in myosin V: from small conformational changes to large directed movements. *PLoS Computational Biology*, 4(8), 2008.

[14] M. Cecchini, P. Kolb, N. Majeux, and A. Caflisch. Automated docking of highly flexible ligands by genetic algorithms: A critical assessment. *Journal of computational chemistry*, 25(3):412–422, 2004.

[15] M. Clark, RDI Cramer, and N. Van Opdenbosch. The tripos force field. *J. Comput. Chem*, 10:982–1012, 1989.

[16] H. Claussen, C. Buning, M. Rarey, and T. Lengauer. FlexE: efficient molecular docking considering protein structure variations1. *Journal of molecular biology*, 308(2):377–395, 2001.

[17] I.W. Davis, A. Leaver-Fay, V.B. Chen, J.N. Block, G.J. Kapral, X. Wang, L.W. Murray, W. Bryan Arendall III, J. Snoeyink, J.S. Richardson, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research*, 2007.

[18] R.S. DeWitte and E.I. Shakhnovich. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc*, 118(47):11733–11744, 1996.

[19] F. Dey and A. Caflisch. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inf. Model*, 48(3):679–690, 2008.

[20] D. Ekonomiuk and A. Caflisch. Activation of the West Nile virus NS3 protease: Molecular dynamics evidence for a conformational selection mechanism. *development*, 4:7.

[21] D. Ekonomiuk, X.C. Su, K. Ozawa, C. Bodenreider, S.P. Lim, G. Otting, D. Huang, and A. Caflisch. Flaviviral protease inhibitors identified by fragment-based library docking into a structure generated by molecular dynamics. *Journal of medicinal chemistry*, 52(15):4860–4868, 2009.

[22] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, and R.P. Mee. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11(5):425–445, 1997.

[23] M. Feher. Consensus scoring for protein-ligand interactions. *Drug discovery today*, 11(9-10):421–428, 2006.

[24] P. Ferrara, H. Gohlke, D.J. Price, G. Klebe, and C.L. Brooks III. Assessing Scoring Functions for Protein- Ligand Interactions. *J. Med. Chem*, 47(12):3032–3047, 2004.

[25] R. Friedman and A. Caflisch. Discovery of plasmepsin inhibitors by fragment-based docking and consensus scoring. *ChemMedChem*, 4(8):1317–1326, 2009.

[26] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem*, 47(7):1739–1749, 2004.

[27] T.M. Frimurer, G.H. Peters, L.F. Iversen, H.S. Andersen, N.P.H. Møller, and O.H. Olsen. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophysical journal*, 84(4):2273–2281, 2003.

[28] E.J. Fuentes, C.J. Der, and A.L. Lee. Ligand-dependent dynamics and intramolecular signaling in a PDZ domain. *Journal of molecular biology*, 335(4):1105–1115, 2004.

[29] V.J. Gillet, W. Newell, P. Mata, G. Myatt, S. Sike, Z. Zsoldos, and A.P. Johnson. SPROUT: Recent developments in the de novo design

of molecules. *Journal of chemical information and computer sciences*, 34(1):207–217, 1994.

[30] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions1. *Journal of Molecular Biology*, 295(2):337–356, 2000.

[31] H. Gohlke, G. Klebe, et al. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, 41(15):2644–2676, 2002.

[32] D.S. Goodsell and A.J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, 8(3):195–202, 2004.

[33] A.P. Graves, R. Brenk, and B.K. Shoichet. Decoys for docking. *J. Med. Chem*, 48(11):3714–3728, 2005.

[34] J. Guenther, A. Bergner, M. Hendlich, and G. Klebe. Utilising structural knowledge in drug design strategies: applications using Relibase+. *Journal of molecular biology*, 326(2):621–636, 2003.

[35] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.

[36] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[37] T. Hansson, C. Oostenbrink, and W.F. van Gunsteren. Molecular dynamics simulations. *Current opinion in structural biology*, 12(2):190–196, 2002.

[38] R.H. Henchman and J.A. McCammon. Extracting hydration sites around proteins from explicit water simulations. *Journal of computational chemistry*, 23(9):861–869, 2002.

[39] A.L. Hopkins, C.R. Groom, and A. Alex. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today*, 9(10):430–431, 2004.

[40] D. Huang and A. Caflisch. Efficient evaluation of binding free energy using continuum electrostatics solvation. *Journal of medicinal chemistry*, 47(23):5791–5797, 2004.

[41] D. Huang and A. Caflisch. Library screening by fragment-based docking. *Journal of molecular recognition: JMR*, 2009.

[42] R.E. Hubbard. 3D structure and the drug-discovery process. *Molecular BioSystems*, 1(5-6):391–406, 2005.

[43] G.T. Ibragimova and R.C. Wade. Importance of explicit salt ions for protein stability in molecular dynamics simulation. *Biophysical journal*, 74(6):2906–2911, 1998.

[44] J.J. Irwin and B.K. Shoichet. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model*, 45(1):177–182, 2005.

[45] G. Jones, P. Willett, and R.C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of molecular biology*, 245(1):43–53, 1995.

[46] W.L. Jorgensen. The many roles of computation in drug discovery. *Science's STKE*, 303(5665):1813, 2004.

[47] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926, 1983.

[48] M. Karplus and J.A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural & Molecular Biology*, 9(9):646–652, 2002.

[49] S. Kazemi, D.M. Krüger, F. Sirockin, and H. Gohlke. Elastic potential grids: accurate and efficient representation of intermolecular interactions for fully flexible docking. *ChemMedChem*, 4(8):1264–1268, 2009.

[50] D. Kern and E.R.P. Zuiderweg. The role of dynamics in allosteric regulation. *Current opinion in structural biology*, 13(6):748–757, 2003.

[51] P. Kolb and A. Caflisch. Automatic and efficient decomposition of two-dimensional structures of small molecules for fragment-based high-throughput docking. *J. Med. Chem*, 49(25):7384–7392, 2006.

[52] P. Kolb, R.S. Ferreira, J.J. Irwin, and B.K. Shoichet. Docking and chemoinformatic screens for new ligands and targets. *Current opinion in biotechnology*, 2009.

[53] P. Kolb, D. Huang, F. Dey, and A. Caflisch. Discovery of kinase inhibitors by high-throughput docking and scoring based on a transferable linear interaction energy model. *Journal of medicinal chemistry*, 51(5):1179–1188, 2008.

[54] DE Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98, 1958.

[55] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, and T.E. Ferrin. A geometric approach to macromolecule-ligand interactions* 1. *Journal of Molecular Biology*, 161(2):269–288, 1982.

[56] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science*, 7:1884–1897, 1998.

[57] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295, 1999.

[58] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, and A. Caflisch. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins Structure Function and Genetics*, 37(1):88–105, 1999.

[59] Y.C. Martin. Let's not forget tautomers. *Journal of Computer-Aided Molecular Design*, 23(10):693–704, 2009.

[60] J.A. McCammon, B.R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[61] G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.F. Truchon, and W.D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model*, 47(4):1504–1519, 2007.

[62] F.A. Momany and R. Rone. Validation of the general purpose QUANTA® 3. 2/CHARMm® force field. *Journal of Computational Chemistry*, 13(7):888–900, 1992.

[63] I. Muegge and Y.C. Martin. A General and Fast Scoring Function for Protein- Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem*, 42(5):791–804, 1999.

[64] C.W. Murray, C.A. Baxter, and A.D. Frenkel. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *Journal of computer-aided molecular design*, 13(6):547–562, 1999.

[65] J. Norberg and L. Nilsson. Advances in biomolecular simulations: methodology and recent applications. *Quarterly Reviews of Biophysics*, 36(03):257–306, 2004.

[66] M. Novinec and A. Baici. Conformational flexibility and allosteric regulation of cathepsin K. 2010.

[67] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, 261(3):470–489, 1996.

[68] M.M. Rhodes, K. Réblová, J. Šponer, and N.G. Walter. Trapped water molecules are essential to structural dynamics and function of a ribozyme. *Proceedings of the National Academy of Sciences*, 103(36):13380, 2006.

[69] V. Schnecke, C.A. Swanson, E.D. Getzoff, J.A. Tainer, and L.A. Kuhn. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins Structure Function and Genetics*, 33(1):74–87, 1998.

[70] T. Schulz-Gasch and M. Stahl. Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discovery Today: Technologies*, 1(3):231–239, 2004.

[71] B.K. Shoichet, A.R. Leach, and I.D. Kuntz. Ligand solvation in molecular docking. *Proteins Structure Function and Genetics*, 34(1):4–16, 1999.

[72] G.M. Sueel, S.W. Lockless, M.A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural & Molecular Biology*, 10(1):59–69, 2002.

[73] J.F. Swain and L.M. Gierasch. The changing landscape of protein allostery. *Current opinion in structural biology*, 16(1):102–108, 2006.

[74] T. ten Brink and T.E. Exner. Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein- Ligand Docking Results. *J. Chem. Inf. Model*, 49(6):1535–1546, 2009.

[75] C.J. Tsai, B. Ma, Y.Y. Sham, S. Kumar, and R. Nussinov. Structured disorder and conformational selection. *Proteins: Structure, Function, and Bioinformatics*, 44(4):418–427, 2001.

[76] ML Verdonk, V. Berdini, MJ Hartshorn, WT Mooij, CW Murray, RD Taylor, and P. Watson. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *Journal of chemical information and computer sciences*, 44(3):793.

[77] M.L. Verdonk, P.N. Mortenson, R.J. Hall, M.J. Hartshorn, and C.W. Murray. Protein- Ligand Docking against Non-Native Protein Conformers. *J. Chem. Inf. Model*, 48(11):2214–2225, 2008.

[78] G.M. Verkhivker, D. Bouzida, D.K. Gehlhaar, P.A. Rejto, S. Arthurs, A.B. Colson, S.T. Freer, V. Larson, B.A. Luty, T. Marrone, et al. Deciphering common failures in molecular docking of ligand-protein complexes. *Journal of Computer-Aided Molecular Design*, 14(8):731–751, 2000.

[79] R. Wang, Y. Gao, and L. Lai. LigBuilder: a multi-purpose program for structure-based drug design. *Journal of molecular modeling*, 6(7):498–516, 2000.

[80] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In *Proc. Edinburgh Math. SOC*, volume 17, pages 1–14, 1970.

Acknowledgements

I consider *life* only as mere a succession of random events and I have no eschatological explanation concerning the stochastic correlations that sometimes happen. But, often, it is very pleasant to have a moment of introspection and retrospectively try to *find* a meaning of correlated stochastic events that populated my life, be they scientific and not.

I heard of the group of **Amedeo Caflisch** from **Fabio Parmeggiani** one day in April (or was it May?) 2005. He was a former master student of my same wonderful supervisor, Prof. **Alejandro Hochkoeppler**, and I was undecided on my future and looking for something to do after my master in industrial chemistry. He made me aware of two things: first of the existence of the city Zürich as a scientific centre of outstanding quality and second of the group of **Amedeo Caflisch**. I would say that he is the first culprit!

I wrote **Amedeo** in the beginning of October 2005, a couple of weeks before my graduation, and he answered (by chance!) just a couple of days after it and invited me for an interview in Zürich in November. Probably, I did a good impression, for he offered me the position. I accepted and left for Switzerland with my paper suitcases on January 13, 2006, and started on January 16. Retrospectively thinking of those four and a half years, I can really say that he has been a very good supervisor. Sometimes, I would have liked to be more pushed and motivated, that is true, but this is probably the only "fault" I could find. I really liked working for and with him, because he is very scientifically stimulating and he always asked for my scientific opinions and ideas, and considered and discussed them, also when I was a *newbie*, and this is highly rewarding (not that now I am the super-expert...). He involved me in very interesting projects – and most of them were *desperate* projects, a politically incorrect synonym for "very difficult and challenging". But I very liked that, truly! I also want to thank P.D. Dr. **Peer Mittl** and Prof. Dr. **Cristina Nevado** for being part of my thesis committee.

Then there are many important people who decorated my Swiss existence. **Beatrice Paoli** was probably the first one, and she offered shelter and protection during the very first two weeks in Switzerland.
I always had good time with her and unfortunately we could form the UFFICIO DELL'AMMMORE only for a few months. Going back to the UFFICIO DELL'AMMMORE topic, another person I would like to introduce is **Marino Convertino**. He was a shy and afraid student from southern Italy, when I had the first contact with him. He called me "Doctor", and I was pleased for that. Eventually we shared altogether the UFFICIO DELL'AMMMORE with **Beatrice**, until she abandoned us. The void left by her absence could not be filled anymore, and the UFFICIO DELL'AMMMORE fast disappeared, as it silently arrived. But fortunately my good friend **Marino** remained!

Then there is **Riccardo Pellarin**, with whom I had one of my bests collaborations. Everything started in high secrecy with the triumvir **Fabio Parmeggiani**, too, and whatever will be of this project, it remains a wonderful experience and I learned a lot, from both. **Riccardo**, in particular, taught me a lot, and I wish him a lot of success in his career. He deserves it and he is really good.

And then there is my good friend **Patricia Schenker**. Without her, my whole doctorate would have been different. As long as she was in the department, she gave me a daily ration of smiles and nice jokes. Not to forget that our collaboration was the first and the most wonderful ever had.

Again, and again on the UFFICIO DELL'AMMMORE topic, I want to say couple of things about **Enrico Guarnera**, because he occupied my desk before my desk was my desk because it was his my desk. He transmitted me (some of) his passion for science, and because of that I started reading about evolution, and therefore I started to realize that living in this world is witnessing to the greatest show ever (adapted from **Richard Dawkins**, an amazing science writer). I hope he will be do well in **New York City, USA**. Reading of science is as wonderful as reading fiction.

And now I can go to the science fiction. I do not know whether I am an **Asimov**-ian or not, but I am very likely to be, for I have read almost everything that is easy obtainable. I like to think that we are somehow similar. He studied chemistry. I studied chemistry. He has a doctorate in biochemistry. I... ehm... not yet. He sucked at research. Me, too. And many other things. I want to remember him because he accompanied me during this last year. Everything started in April 2009 with a copy of "Foundation and Earth" of my father that I devoured. Shortly afterwards, I got almost everything[1].

If I liked this period in Switzerland a lot, it is due to many persons, whose life intersected mine. **Flavio**, **Elisabeth**, and **Alessandro**. Also **Andrea Prunotto**, for I liked very much to play with him and he always had weird problems to solve with programming. That has been very stimulating! Then there is Prof. **Antonio Baici**. He has been very kind to me and his scientific contribute has been essential. His project on Cathepsin B, my very first one, gave me the pleasure of trying *desperate* projects (desperate again in the sense that they are very likely to fail). Having **my violin** repaired by him twice, looking at him while working and chatting about violin and music related things has been really wonderful. All my colleagues and co-workers were also important for my stay in Zürich and for this thesis. Thank you all! Scientifically, I want to thank **Peter Kolb** (my "mentor"), great in science and also a gentleman and **Francois Marchand**, or Mar-Charmm! Thanks! And **Sandra Rennebaum** expecially for a translation! **Marko Novinec** because of having involved me in the very interesting field of allostery and because of all the experimental work. **Christina Ewald** and **Gautham Varadamsetty** because of their very good experiments in the Armadillo project.

[1] My girlfriend **hates** him...

Vorrei ringraziare tutta la mia famiglia e i miei parenti. I miei genitori per avermi cresciuto bene, per avermi dato amore e avermi stimolato. In particolare **mio padre Giuseppe** per avermi trasmesso la passione per i computer. Chissà dove sarei adesso se non mi fossero mai piaciuti! **Mio fratello Cosimo**, un costante aiuto con la programmazione in questi anni. Fortunatamente ad entrambi piace Python! **Mia madre Eva**, anche, ovviamente, ma purtroppo non ha un ruolo diretto in questa tesi di dottorato. Però pensandoci, se non sono morto di fame è perché mi ha insegnato in breve tempo, prima della mia partenza per Zurigo, quei quattro rudimenti di cucina. Poi, il fatto che io sia un fenomeno in cucina è ovviamente dovuto alla mia formazione di chimico... :-) **Sofie,** Sofie, **Sofie**. Ripeterei il tuo nome mille e mille volte! (ma forse già lo faccio, no?). Senza di te sarebbe stato tutto diverso, e sarà sicuramente migliore una volta finito tutto questo! Grazie!

Wie der Behemot die Meere durchstürmt, so durchflog er die Grenzen seiner Kunst.Vom Girren der Taube bis zum Rollen des Donners, von der spitzfindigsten Verwebung eigensinniger Kunstmittel bis zu dem furchtbaren Punkt, wo das Gebildete übergeht in die regellose Willkür streitender Naturgewalten, alles hatte er durchmessen, alles erfaßt. Der nach ihm kommt, wird nicht fortsetzen, er wird anfangen müssen, denn sein Vorgänger hörte nur auf, wo die Kunst aufhört.

— **Danke.**

Curriculum Vitae

Personal Data

Surname	ALFARANO
Name	Pietro
Date of birth	May 12, 1981
Nationality	Italy

Education

1/2006–now	PhD in Biochemistry University of Zurich
10/2003–10/21/2005	MSc in Industrial Chemistry University of Bologna (Italy) Title: "Factors determining the hypersynthesis of murine interferon α_1 in *Escherichia coli*." Supervisor: Prof. Dr. Alejandro Hochkoeppler Grade: 110/110 *cum laude*.
10/2000–6/18/2003	BSc in Industrial Chemistry University of Bologna (Italy) Title: "New synthesis pathways for light olefins: the oxydehydrogenation of propane to propylene". Supervisor: Prof. Dr. Fabrizio Cavani Grade: 110/110 *cum laude*.
6/2000	High school Scientific Degree Liceo scientifico statale "A. B. Sabin" Bologna (Italy) Grade: 100/100

Publications during master studies

A. Stefan, P. Alfarano, D. Merulla, P. Mattana, E. Rolli, P. Mangino, L. Masotti, and A. Hochkoeppler. The regulatory elements of araBAD operon, cotrary to lac-based expression systems, afford hypersynthesis of murine, and human interferons in Escherichia coli. Biotechnology Progress, 2009.

Publications during doctoral studies

P. Schenker, P. Alfarano, P. Kolb, A. Caflisch, and A. Baici. A double-headed cathepsin B inhibitor devoid of warhead. Protein Science, 17(12):2145, 2008.

V. Exner, E. Aichinger, H, Shu, T. Wildhaber, P. Alfarano, A. Caflisch, W. Gruissem, C. Koehler and L. Hennig. The chromodomain of LIKE HETEROCHROMATIN PROTEIN 1 is essential for H3K27me3 binding and function during Arabidopsis development PLoS ONE, 4(4), 2009.

Congresses attendance

Virtual Discovery, London 2007. Poster.
Computer-aided drug-design, Tilton (NH, USA) 2007. Poster.

Zurich, May 20, 2010
Pietro Alfarano